Fundamentals of EXPERIMENTAL PSYCHOLOGY

2nd edition

Fundamentals of EXPERIMENTAL PSYCHOLOGY

CHARLES L. SHERIDAN
University of Missouri
and Veterans Administration Hospital
Kansas City, Missouri

HOLT, RINEHART AND WINSTON
New York Chicago San Francisco Atlanta
Dallas Montreal Toronto London Sydney

The author wishes to thank the following publishers and authors for permission to reproduce illustrations:

Academic Press, Inc., for figure from R. Walk, A study in visual depth and distance perception in animals. In D. S. Lehrman, R. A. Hinde, and E. Shaw. *Advances in the study of behavior,* 1965.

The American Psychological Association for figures from the following: From D. D. Wickens, H. M. Schroder, and J. D. Snide, Primary stimulus generalization of the GSR under two conditions, *Journal of Experimental Psychology,* 1954, *47,* fig. 2, p. 53. Copyright 1954 by the American Psychological Association. From J. A. Swets, W. P. Tanner, and T. G. Birdsall, Decision processes in perception, *Psychological Review,* 1961, *68,* fig. 3. Copyright 1961 by the American Psychological Association. From E. Hilgard, Pain as a puzzle for psychology and physiology, *American Psychologist,* 1969, *24,* fig. 7, p. 109, and fig. 8, p. 100. Copyright 1969 by the American Psychological Associaton. From S. Milgram, Group pressure and action against a person, *Journal of Abnormal and Social Psychology,* 1964, *69,* fig. 1., p. 140. Copyright 1964 by the American Psychological Association. From S. Milgram, Liberating effects of group pressure, *Journal of Personality and Social Psychology,* 1965, *1,* 127–143. Copyright 1965 by the American Psychological Association. Excerpt from *Ethical principles in the conduct of research with human participants.* Copyright 1973 by the American Psychological Association. Reprinted by permission.

Appleton-Century-Crofts for figures from *Great experiments in psychology* by H. E. Garrett, 1951.

Dover Publications, Inc., for figure from *Principles of animal psychology* by N. R. F. Maier and T. C. Schneirla. Dover Publications, Inc. New York, 1964. Reprinted through permission of the publisher.

W. H. Freeman and Company for figures from the following: From Experiments in discrimination by N. Guttmann and H. I. Kalish. Copyright © 1958 by Scientific American, Inc. All rights reserved. From Plasticity in sensory-motor system by R. Held. Copyright © 1965 by Scientific American, Inc. All rights reserved.

Holt, Rinehart and Winston for figures from the following: From *Psychology,* Revised Edition, by Delos D. Wickens and Donald R. Meyer. Copyright © 1961 by Holt, Rinehart and Winston, Inc. From *Animal structure and function* by Donald R. Griffin. Copyright © 1962 by Holt, Rinehart and Winston, Inc. From An essay on dreams by William C. Dement from *New directions in psychology II* by Frank Barron et al. Copyright © 1965 by Holt, Rinehart and Winston, Inc. Reproduced by permission.

The Humanities Press Inc. for figure from K. Gottschaldt, The influence of past experience on the perception of figures (1926). In *A source book of gestalt psychology,* W. D. Ellis (Ed.), 1938.

International Publishers for figure from *Lectures on conditioned reflexes* by I. P. Pavlov, 1928.

The Journal Press for figures from the following: From The mechanism of vision: I. A method for rapid analysis of pattern-vision in rats, by K. S. Lashley, *The Journal of Genetic Psychology,* 1930, *37,* fig. 1, p. 454, and fig. 2, p. 456. From The generalization of conditioned responses: I. The sensory generalization of conditioned responses with varying frequencies of tone, by C. I. Hovland, *The Journal of General Psychology,* 1937, *17,* fig. 2, p. 136.

McGraw-Hill Book Company for figures from the following: From *Basic psychology,* by R. Stagner and C. M. Solley, 1970. From *Psychometric methods* by J. P. Guilford. Copyright 1954, McGraw-Hill Book Company. Used with permission of the McGraw-Hill Book Company.

The M.I.T. Press for figure reprinted from S. S. Stevens. The psychophysics of sensory functions. In W. A. Rosenblith, *Sensory communication* by permission of The M.I.T. Press, Cambridge, Massachusetts.

Routledge & Kegan Paul Ltd. for figure from K. Gottschaldt, The influence of past experience on perception of figures (1926). In *A source book of gestalt psychology,* W. D. Ellis (Ed.), 1938.

John Wiley & Sons, Inc., for figures from the following: From *Applied experimental psychology* by A. Chapanis, W. R. Garner, and C. T. Morgan, 1949. From R. Fantz, Visual perception and experience in early infancy. In H. W. Stevenson's *Early behavior,* 1967. From R. M. Dawes, *Fundamentals of attitude management,* 1973. Reprinted by permission of John Wiley & Sons, Inc.

Preface to First Edition

Experimental psychology comprises at least four major areas: strategy, tactics, heuristics, and logistics. A complete textbook of experimental psychology should give its reader adequate preparation in all four areas. Strategy involves the broadest definition of goals and methods. Issues that are normally regarded as philosophical, but which have pragmatic importance to the experimentalist; methodological questions such as those having to do with statistical versus nonstatistical procedures, behaviorism versus phenomenology; or any other issue whose solution has highly general import should be included under strategy.

An "armchair" psychologist may be satisfied with only an understanding of strategy, but the rough-and-ready experimentalist prefers direct investigation and tends to leave strategic questions to be answered by someone else or to depend on his own intuitive solutions. Experimentalists usually "muddle through" or "follow their noses" in designing and conducting experiments. Thus there exists a considerable amount of procedural information that is less general in its application than strategic information. This sort of information I consider to be tactical, and since it has not been codified into a set of rules it can best be incorporated into a text by describing the specific problems of interest to experimentalists and the experimental devices that have been used to solve them. The best way to learn tactics is by doing experiments under the tutelage of a master experimenter. That sort of experience is not likely to be supplanted by any text, but the text can help to extend the range of an apprentice experimenter's experience and will supplement his laboratory work, present or future. In addition to helping him devise useful "rules of thumb," tactical descriptions of experiments should give him an idea of which research areas have been and currently are of interest to experimental psychologists. Within the framework of a broad strategy, many different areas of interest have been pursued by experimental psychologists. It is to these areas that the novice experimentalist should turn his attention, for he needs the stimulation of grappling with problems that have fascinated his predecessors, and his study of their methods prepares him for devising new approaches.

Heuristics is the discipline that attempts to formulate principles of discovery. Which methods will or will not lead successfully to an investigative technique of science? Unfortunately, there is little to be said with confidence about heuristics, since we have no science for formulating science. There are common arguments, however, whether one should or should not use theory as a guide to research. Undoubtedly the novice experimentalist should learn about such arguments, but there is not enough solid information to justify inclusion of a complete section in the present text. My approach has been to scatter heuristic issues throughout the book, wherever they seemed to blend in with specific subject matter. Even though a section on heuristics does belong in a textbook of experimental psychology, heuristic principles must be more firmly established before this inclusion can be made.

Logistics presents a similar problem. Logistics has to do with the practical problems involved in implementing an experiment. What kind of apparatus will do the job? Where can it be obtained? What electronic circuits will accomplish the task? Even the question, "What do the stripes on a resistor mean?" is a logistic problem. It would be possible to treat logistics in an introductory textbook, but I have not done so because too much of the information is ephemeral, because it would make the book too long and cumbersome, and because even massive efforts by others to present logistics as a text entity might result in desultory discussion without helpful conclusions. This is unfortunate, since my experience in training students to do research leads me to suspect that the inability to cope with simple logistic matters (even to make judiciously persistent phone calls) is a major obstacle for most students. Just as I have assigned heuristic questions to various nooks and crannies of the text, I have injected a number of logistic matters from time to time. For example, many commonly used types of apparatus are described, and for the most part these are a concomitant of descriptions of experiments.

In sum, then, the textbook is divided into a section on strategy and a section on tactics, with heuristic and logistic issues interspersed.

The prospective user of this book has a right to know certain attitudes that have influenced my mode of presenting experimental psychology. One of them is that the questions about "mind," which encourage some students to select psychology courses, should be taken seriously. It is common for psychologists to joke or feel guilty about the fact that students study psychology in order to find out about "sex, hypnosis, and personality," but that instead we teach them about optimal CS-UCS intervals, intermittent reinforcement, and the variables influencing the rate of learning lists of nonsense syllables. After the first few weeks of indoctrination, a student's use of the word "mind" is cause for classroom tittering. My attitude toward such questions is that they should not be rejected, but analyzed. Questions

framed in ordinary language have many facets. There are really many different questions, to some of which we answer "yes," to some of which we answer "no," and to others of which our answer is that the question is "meaningless." I have great respect for ordinary language, and even suspect that its frequent ambiguity is not without behavioral function. My tendency is to begin with the questions of ordinary language and to make a gradual transition to the meanings that the original terms have come to have in scientific discourse. Thus, for example, experimental psychology is defined as the experimental study of mind, and then the reasons for developing an objective psychology are provided so that the special scientific meanings of "mind" can be understood.

A related issue is that of relevance. It is my opinion that experimental psychology is vitally relevant to everyday problems of human life. As a student, in my first encounters with psychology it seemed as though psychologists were, for mysterious reasons of their own, interested in the salivary responses of dogs and in the subtleties of behavior displayed by a special breed of rat in oddly restrictive situations. It was as though a strange breed of men had been born with the affliction of being psychologists and with bizarre interests that were manifestations of the affliction. It seems to me that students should be shown the tortuous path than can lead a reasonably talented literary man to give up writing in order to dedicate his life to the study of "delay of reinforcement" or some other seemingly reprehensible topic. In this book I have attempted to show how grenerally applicable are the methods of experimental psychology and how generally applicable are the principles derived from experiments done on salivating dogs and key-pecking pigeons. Put simply, I have attempted, where possible, to place experimental psychology clearly within the context of our culture, both by emphasizing the potential relevance of basic material and by including a short chapter on the application of experimental psychology to the study of complex human behavior, particularly by experimental social psychologists.

My third attitude may seem more like a "caveat." It is that students should be given the whys and wherefores, and the pros and cons, for the various theoretical and methodological positions presented. As much as possible they should be given the opportunity to make an informed decision of their own about their acceptance of contending positions. A consequence of this is that, despite all efforts at clarity, the going may be rough at times. But a superficial survey that offers students mere catchwords ("What is gestalt psychology?" "The whole is more than the sum of its parts.") without providing understanding strikes me as the very nemesis of scholarship. I have tried to avoid losing touch with the student reader while providing him with more than the usual depth of understanding of the issues. To some extent my

own students, who have gone over much of this material, have shown that the task can be accomplished. But it is difficult, and only experience will tell whether others will have similar success.

The policy of presenting pros and cons for areas of dispute extended into the areas of data reliability and generality, and the end product requires a word of explanation. There is no section on statistics in this book. Originally, a brief, practical summary of statistics was contemplated, but several readers of the early manuscript plan argued against its inclusion because they thought that it could not possibly clarify statistics sufficiently for the unprepared student and would have little value for the student familiar with statistics. For this reason, I deleted the section on statistics, even though the questions of datum reliability and generality are vitally important strategic issues. Statistical methods are commonly used to evaluate the reliability and generality of data, but these methods are not the only ones available. In fact, there is a great deal of contention among contemporary psychologists over whether statistical or "experimental" analyses of behavior are to be preferred. I have tried to develop a general treatment covering the issues of datum reliability and generality—a treatment that encompasses both the statistical and experimental analyses of behavior. This approach reflects the actual practices of experimental psychologists rather than any theoretical design or personal preference for methods of evaluating data.

Some readers may be puzzled at my selection of research materials for presentation. It is important to realize that no attempt has been made to give a truly comprehensive coverage of all important work in experimental psychology. This strikes me as patently impossible in an introductory text. The selected research is intended to introduce the reader to interesting and important research tactics. Hopefully, many of these tactical devices have a generality that should be useful to those whose research interests run to somewhat different topics. Since the selections are included in order to illustrate tactical principles, they should not be read as though they were literature reviews. A great deal of valuable research (including most of my own) that might have been included has been omitted, but not because it is considered unimportant. Put simply, the material included is tactically valuable, interests me, and has worked well with my students.

It is customary for authors to emphasize the great amount of help they received from others in the course of writing their books. In the process of writing this one, I have learned that such proclamations are not mere rhetoric. I am indebted to a great many people. Very early in the development of the manuscript, I received a great deal of helpful encouragement from Frances J. Temple, from William A. Hillix, and from Ralph N. Haber. Without this encouragement, I doubt whether I could have made the sacrifices necessary for completion of the book.

Many people contributed by making critical comments and giving constructive advice on the basis of earlier versions of the manuscript. The most important critical contribution was made by Kendra S. Vaughan, who read the manuscript from the point of view of an "intelligent layman" and who goaded me to clarify many points that might be obscure or difficult for a novice to understand. She also contributed to typing a large part of the manuscript. I am also indebted to Betty Stevenson and the Research Services section of the Graduate School at the University of Missouri at Kansas City for making a major contribution to the preparation of the manuscript. Manuscript preparation was also greatly facilitated in a variety of ways by Richard King, my graduate assistant, who volunteered many hours needed for the final steps of preparation. Kim Groves also carried a heavy share of this final burden. With respect to manuscript preparation, I must also express my gratitude to Cheryl Salter and Ricki Bronfman, without whom the typing would have been seriously delayed.

Diane D. Edwards, with whom I have "team taught" experimental psychology over the past few years, has influenced greatly my extensive examination of many issues related to her main area of interest—operant conditioning. If the book proves to be at all palatable to those enthusiastic about operant conditioning, it will be largely due to her efforts.

Four of my students made scholarly contributions of noteworthy value to the book. The first of these was Robert E. Sanders, who was my major intellectual stimulus during the bulk of the writing. It would be difficult to identify his specific contributions, since we contemplated, meditated, speculated, and cogitated so much together that our thinking often melded into a unity. Earle G. Wallingford brought to my attention a great deal of the literature on microelectrode studies of color vision, and both Leigh Ruhlen and Virginia Wolfe contributed to my knowledge of the literature on pain.

Finally, I want to thank Don R. Justesen, Chief of Neuropsychology at the Kansas City Veterans Administration Hospital, whose kindness, flexibility, and helpfulness have stimulated my productivity over the past several years.

C. L. S.

Kansas City, Missouri
January 1971

Preface to Second Edition

I have made a number of major changes in this second edition of *Fundamentals of Experimental Psychology.* The most important one has been what I hope is a marked increase in readability. I have gone to great lengths to simplify the language. The present edition covers the needed material but in a much easier style.

I have also added a number of aids to the student and teacher. There are now study questions at the end of each chapter. The key ideas of the book have been placed in boxes. These boxes are meant to be used in two different ways. They may serve as review material, but they are also designed to be read independently, as an alternative to reading the full material. Teachers may not have the time or the inclination to cover some of the sections in full detail. They can give their students a broad overview by having them read Key Ideas Boxes. This, hopefully, will provide maximum flexibility in using the book.

A second type of box, called a Logistics Box, has been used to lay out supportive material that would otherwise break the continuity of the narrative. These boxes occur frequently in the section on the design of experiments, where they are used to provide worked examples of statistical analyses.

Another important addition has been the incorporation of exercises, either at the ends of chapters or in the body of the text. These exercises can be done with equipment readily available at home, or at least with equipment that is very inexpensive. The exercises are actually enough to serve as a laboratory manual. Where there is no full-fledged laboratory, the interested students can do them on their own.

It has been an important experience to teach from the first edition, then deal with the students at later stages in their training. It seemed to me that there were many practical things more important than some of the more ethereal ones I had emphasized in the past. Though it is true that no book can really make a person into an experimental psychologist, one can help. I have tried to make the present edition as practical as possible. For example, I have added worked examples of such things as statistical and psychophysical procedures. With my own students, I have found that little teaching needs to be done other than to refer them to the worked examples in these cases. A variety of other practical matters have been incorporated. Guidance is given in how to get ideas for research, how to read the research literature, and how to present the results of research either orally or in writing. In a nutshell,

I have tried to deal with both the formal and the informal aspects of experimental methodology.

I have expanded the treatment of human behavior. This occurs in many places in the text. The most noteworthy expansions in this regard have been the addition of a section on human information processing, a section on dealing with human subjects (ethics, role playing versus deception, demand characteristics), and a treatment of the problems entailed in the use of verbal reports.

It might be of value for me to explain my treatment of selected research. It is easy to get the impression that the chapters on selected research are meant to be reviews of the literature. This is not the case. My view is that, though on the one hand there are certain aspects of experimental psychology that have been systematized and can be taught as a set of principles and rules (experimental designs fall into this category), many other aspects of experimental psychology have not been systematized. Wherever possible, I have tried to systematize them. But, to some degree students need to see examples of experimental psychologists at work. In this way they have an opportunity to get a feel for the nuances of what I have called the tactics of experimental psychology. I am a firm believer in the Kantian dictum that "Percepts without concepts are meaningless; concepts without percepts are empty." Students need sample experiments (and laboratory experiences!) to flesh out the bare bones of systematized experimental methodology. But the intent here is nevertheless methodological, though in a broader than usual sense—one that includes informal aspects of method. Thus the chapters on selected research are not reviews of the literature, though they are partially that. Nor are they mere recitations of examples of the designs presented in earlier chapters, though they are partially that too. Perhaps it is best to say that they are vignettes of psychologists in action, along with some methodological commentary.

Finally, let me say that I have not forgotten that vast bulk of students who will never be experimental psychologists, but who take courses in experimental psychology nevertheless. In fact, most of my own students fall into this category, though many of them end up experimental psychologists in spite of their original intentions. I feel very strongly that the course in experimental psychology should not be a meaningless exercise for students of this type. Though they will not be producers of research, they will most certainly be consumers of it. In newspapers, magazines, on television, virtually everywhere we turn there is a barrage of information or misinformation concerning psychology. It is a worthwhile exercise to employ a part of one's education getting an inside glimpse of where it all comes from. Most importantly, it is of value to learn to be critical about which claims are or are not valid. I hope this book will give the research-consumers that glimpse and teach them how to recognize valid psychological claims. And I hope it will do so relatively painlessly.

Many people have helped in one way or another in making this book into a reality. I will not try to acknowledge all of them, but three people were especially important. Deborah Doty of Holt, Rinehart and Winston took a very active role in fostering the book and, in particular, had a lot to do with the increased readability of the second edition. Mickey Ray suffered through taking care of many details essential to the book. Without her it would have been much delayed. And Peggy Sheridan tolerated me during the final frenetic period of production and even returned to her long-abandoned schoolmarmish ways long enough to give the proofs a careful, if rushed, reading.

C. L. S.

Kansas City, Missouri
January 1976

Contents

1 EXPERIMENTAL PSYCHOLOGY AND ITS METHOD

The nature and objectives of experimental psychology

I can't tell you how many times I've been through this. I'm at a party. I meet someone (call her Millie G.) and make light conversation for a little while. People usually ask me what I do for a living. I think it's because I look like a truck-driver but talk like a professor. Hesitantly (will she go away?) I answer, "I'm a psychologist."

Millie's face gets a little more serious, "Oh. That sounds interesting." I can see she's wondering if I can detect her inner secrets. "But I'm not the kind of psychologist you think," I add hastily. "I don't treat patients. I do research. I'm an *experimental* psychologist."

Her face brightens. "Oh, right! My roommate took experimental psychology. That's where you play around with rats!"

It's enough to make a person write a book.

Experimental psychology is a comprehensive discipline. Rats, mazes, nonsense syllables are all things that experimental psychologists use *sometimes.* And they can all be used to good ends. But if you think they are the heart of the matter, you are missing the essence of experimental psychology.

Experimental psychology is simply the use of the methods of experimental science for the purpose of investigating psychological matters. The users of methods of experimental science can raise and answer certain questions about hatred, memory, Zen meditation, vision, mutual attraction, decision making, or any other psychological matter.

Experimental psychologists have helped to create the impression that their field is a narrow one. For instance, *The Journal of Experimental Psychology* does not often have articles on experimental clinical psychology, experimental social psychology, or even physiological psychology. But such traditions are just accidents of history. Like most scientific disciplines, experimental psychology started out with emphasis on the simplest of its potential topics. These came to be identified with the field as a whole. Only later was the full range of experimental psychology made apparent.

Whenever experimental method is applied to psychological events, it is an instance of experimental psychology. It's just that simple.

IMPLICATIONS OF THE DEFINITION OF EXPERIMENTAL PSYCHOLOGY AS METHOD

Basic experimental science often produces information of great practical value. An understanding of the laws underlying a phenomenon is worthwhile in its own right. But there is great advantage in recognizing the direct applicability of experimental method to complex, socially meaningful subject matter. My belief is that psychologists should be trained to recognize that their questions, socially applicable or otherwise, can generally be handled within the framework of experimental method.

Furthermore, it is only logical to define experimental psychology to include the widest possible range of topic areas. There is no rational basis for excluding those instances in which experimental method is applied to complex clinical and social subject matter. Experimental psychology is not realistically represented today by the old narrow definition. The fact is that experimentalists *do* carry out research on complex psychological phenomena, such as conformity, attraction between persons, and mental illness.

METHODOLOGICAL SUBDIVISIONS OF EXPERIMENTAL PSYCHOLOGY

Customarily experimental psychology has been divided according to its various areas of content. These are such topics as Sensation and Perception, Learning, Memory, and Attention. But since experimental psychology is methodological, I divide it according to the various

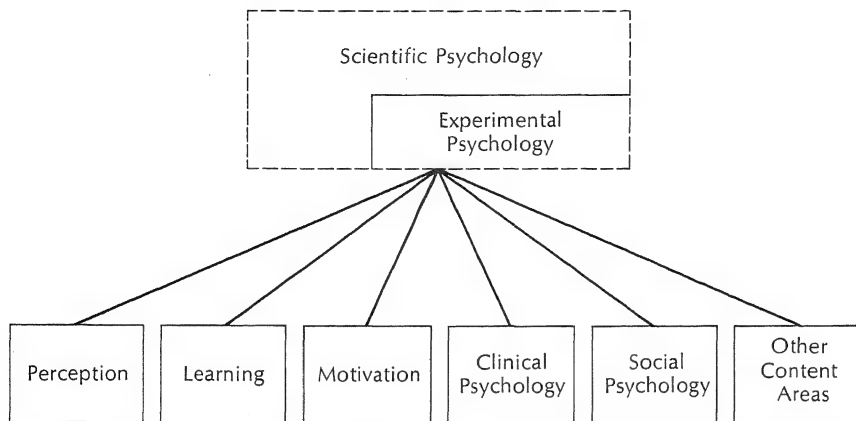



FIGURE 1.1 Experimental psychology is a part of scientific psychology (which includes nonexperimental techniques such as those of naturalistic observation). Experimental psychology has application in all the content areas of psychology. It is a method rather than a content area.

phases of its method. They are Strategy, Heuristics, Logistics, Tactics, and Communication. Method is primary and content is secondary (see Figure 1.1).

Strategy encompasses the selection of goals to be reached, and the selection of the general methods to be used in reaching the goals. For example, some psychologists have selected for themselves the goal of "predicting and controlling human behavior." Others want to "understand human experience." Some psychologists choose to use the experimental method. Others do not. Some psychologists use statistics, others feel that they are the Devil's curse. All of these choices are strategic. Often we make them half-consciously, taking for granted the tradition in which we got our training. It might be better if we chose on more explicitly rational grounds.

Heuristics is concerned with the art and science of discovery. Experimenters can simply depend on their own creative impulses as sources of good research ideas, but certain principles can extend their creative powers. These are included in the coverage of heuristics in Chapter 14.

Logistics is the problem of supply. What kind of apparatus will be used in a given experiment? Where can experimental subjects be obtained? What limitations are imposed by ethical and legal restraints? What kind of psychological test will measure the psychological process to be studied?

In *tactics* we specify the maneuvers to be used in a particular situation in order to reach the objectives specified by strategy. Say you want to know why people come to be attracted to each other. Suppose you have made a strategic decision to rely on verbal reports. You must

devise a particular interview technique or, perhaps, a particular questionnaire. Suppose you have made a strategic decision to rely on behavioral phenomena. You must arrange a tactic for observing pairing. You might arrange meetings in the waiting room of your laboratory during which some measure of attraction can be observed systematically.

Another experimenter who wants to know whether children must learn to see, or whether they are born with this ability, must devise a tactic for measuring their perception without requiring them to report it verbally; and he must also measure it within the very modest level of movement displayed by human newborns.

Communication includes any of the ways we convey the results of experiments to others. Oral presentations (usually at meetings of professional societies) and written publications are the major forms of communication. They are an essential part of the process of science.

THE OBJECTIVES OF EXPERIMENTAL PSYCHOLOGY

Experimental psychology has three main objectives (see Figure 1.2). The first objective is to *measure* psychological events. The second is to *specify* and *predict* the conditions under which given psychological events will occur. The third is to become able to *produce* a desired psychological event at will, given adequate physical and technological support.

These are merely the general objectives of experimental science, adapted to psychology. Scientists *measure* phenomena, they *analyze* them, and they *synthesize* them. Measurement helps greatly in the process of analysis. Analysis helps us discover basic principles or laws. These laws eventually allow us to write equations (or make other models) that predict events not yet observed. Synthesis means reproducing something after the laws underlying its occurrence are known. Synthesis benefits us in two ways: by assuring us of the accuracy and completeness of our understanding of the phenomenon under study, and by working out how to put our knowledge to practical use.

Science and its alternatives

DIFFICULTY OF APPLYING SCIENCE TO PSYCHOLOGY

Teachers of scientific psychology are often in a predicament quite different from anything faced by other science teachers. When people enroll in courses in physics, chemistry, or astronomy, they are ready to

Key Ideas Box 1.1: Experimental psychology

Experimental psychology encompasses all applications of scientific experimental method to psychological events. In principle, its subject matter includes all topic areas of psychology. It is distinct from other approaches because of its experimental *method,* not because of its *content.*

The main methodological subdivisions of experimental psychology may be viewed as *strategy, heuristics, logistics, tactics,* and *communication.* Strategy involves identification of *goals* and main *methods* to be used to reach them. Heuristics involves principles of *discovery,* getting creative research ideas. It is a formulation of and an attempt to aid natural creativity. Logistics involves all matters of *supply.* Such issues as sources of participants and measuring devices typify logistics. Tactics involves the creation of specific methods for *implementing* strategies. Communication has to do with letting others, especially the scientific community, know of your results, *publishing* them.

The objectives of experimental psychology are to *measure* psychological events, to *specify the conditions* under which they will occur, and to *produce* them. This is simply the application of the customary scientific procedures of measuring, analyzing, and synthesizing the events under study.

assume that there are natural laws underlying the behavior of things. Even with the current upsurge of interest in the occult, few students believe that the planets move around because each is propelled by an angel. This is a viewpoint held by certain medieval thinkers. An astronomy teacher would be astonished if a student interrupted a lecture to argue that the heavenly bodies move because of the pull of love ("The love which moves the sun and the other stars"), as the poet Dante said. Modern people are generally willing to assume that the universe follows certain natural laws.

But many people feel that psychological events are an exception to the general rule of lawfulness in nature. The mind, they feel, is free. It cannot be grasped by the objective methods of science. And even if it could be so contained, who would want to reduce the spontaneity of human beings to mere objective laws?

As scientific psychologists face such doubts today, other kinds of scientists have faced them in the past. Remember, Galileo (1564–1642) was condemned and humiliated for saying that the earth moved around the sun. People in authority believed that the earth *had* to be the

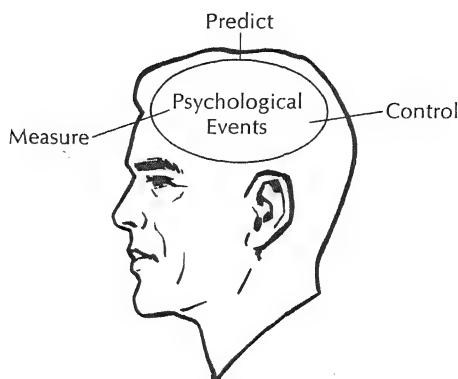



FIGURE 1.2 Objectives of experimental psychology.

unmoving center of the universe. This must have seemed intuitively obvious to them. It seemed well established by traditional writings, and it seemed essential to the notion that man is the pinnacle of creation. But, in deference to well-founded facts, we have given up that view. We are now comfortable with the realization of our very humble position as a speck nowhere near the center of our galaxy, wherever that may be in the universe.

Scientific method, whenever it is introduced into a new sphere of human thought, meets strong resistance. This is because of its unfamiliar and seemingly destructive approach to old questions. But unfamiliarity is not a criterion for rejection. Repeatedly, people have jumped to the conclusion that scientific progress would destroy some traditional value, only to decide 50 or 100 years later that the apparent incompatibilities did not exist. If we wish to maintain objectivity, we must judge new approaches on their own merit, as dispassionately as possible.

Few people are as much in the dark today as those of earlier times who condemned the sun-centered view of the universe, confident that they were doing the Lord's work. Yet we run the risk of similar confusion if we reject objective psychology without demanding of ourselves a well-reasoned basis for the rejection.

You should realize that you probably tend to begin your study of experimental psychology with a learned bias against many of its assumptions. You may also have important misconceptions concerning their significance. If you want to gain an accurate, balanced view, you should treat your "common-sense" attitudes with a little more suspicion that you normally would.

To summarize the point, then, the reader is asked to allow for a bit of unfamiliarity in the argument to follow. You are asked to keep your mind open to what is being proposed, despite its unfamiliarity. You are not expected to accept it all uncritically. Perhaps the best way to emphasize the point is by borrowing a quotation from Descartes's

Discourse on Method. It's aptness has previously been recognized by Keller and Schonfeld (1950). "In order to attain truth, it is necessary once in a man's life that he put aside all the opinions which he has held in the past, and rebuild anew, from the very foundations, his entire system of knowledge." (Loose translation by the author.)

SCIENCE, MEASUREMENT, AND RISKY PREDICTION

Now we can proceed to ask ourselves more specifically what it is that experimental psychologists are trying to accomplish. Their enterprise began in the middle of the nineteenth century. This was when physicists and physiologists were starting to apply their familiar techniques to the study of psychology. Science had been eminently successful in its other areas of application. So why not try it in the mental sphere? This is not so easy as it sounds. It requires that we identify the hard core of scientific method—what is essential to it. We need to distinguish that core from the nonessential attributes that stem from the application of science to a particular subject matter. The mind cannot be weighed in an analytical balance. Nor can it be burned to find how many calories it yields. But does this mean that it cannot be dealt with scientifically?

Some people insist that psychology, if it claims scientific status, must slavishly imitate the methods of physics. Such an approach is called *physicism.* But some of the characteristic methods of physics are probably unique to its subject matter and not essential to a science as such. To restrict the designation "science" to such a close imitation of physics would be a mistake. On the other hand, we should not use the term "science" too loosely. Some people have characterized science by its *empiricism.* Empiricists say that experience, especially that of the senses, is the only source of knowledge. But even something as nonscientific as a poem or a novel is usually based on the experience of the writer. Even the very opposite of physicism, mysticism, uses a kind of experience. Thus, personal knowledge is not the crux of science.

If empiricism fails to discriminate "science," what does? Actually, scientific method can be characterized in many different ways. For our purposes, its most important characteristic is the demand for measurement. The lowest kind of measurement is simple agreement between observers about the presence or absence of something (interobserver agreement) and, of course, consistency from one time to another in the same observer (intraobserver reliability). In science there are often disagreements about how to *interpret* observations, but rarely do we have disagreement about the observed facts themselves. In science, observers agree on the facts unless someone is in error. It may be necessary to instruct observers in order to get high levels of agreement.

Some have identified science by the tendency of scientific theories to

generate *risky predictions.* It might seem that the best theory would be one that explains every possible event that might occur. But such a theory would be useless in helping us to predict. Imagine a theory that could only let us predict that coin-tossing would result in either heads, tails, or balancing on edge. Assuming that these are all the possible outcomes, it would be no better than one that said "We cannot predict which of the possible outcomes will actually occur." To predict everything is no better than to predict nothing. A scientific theory must enable us to make predictions that can prove wrong, depending on the outcomes of observations in the real world.

Risky prediction (which places emphasis on theory) and inter- and intraobserver reliability (which emphasizes fact) have much in common. A prediction is risky if observers can reliably agree on an observation or fact that will refute the theory. Agreement among observers is the backbone of science. This is what is meant when we hear it said paradoxically that scientific knowledge, the most objective form of human knowledge, is social in nature. It is objective precisely because observers can agree on the facts—indeed, this is the very meaning of "objective."

Acceptable degrees of reliability

Granting that the criterion of inter- and intraobserver reliability constitutes the heart of science, we may ask just how great the reliability must be before a discipline can be legitimately considered scientific. Must agreement be perfect, or is some high but less than perfect level of agreement acceptable? Who must agree? Must every human being who has sight and speech agree? Does this include psychotics? The problems raised here are formidable. It is often stated that two or more observers must agree as to whether a thing is present or absent, but this is not entirely clear, nor does it reflect the actual practice of scientists. It is doubtful whether the scientific community would be convinced by two inmates of a mental institution, though they constitute the minimal two observers. Since a rift between conception and practice tends ultimately to result in confusion, it is important that we realize that *no formal criterion of observational agreement is used by scientific investigators,* nor does it seem likely that such a criterion is feasible. Most students learn in their introductory philosophy course that it is almost impossible to generate a completely acceptable definition of such commonly used terms as "justice," "good," or even "table" or "chair." Yet we seldom encounter any practical misunderstandings because of our lack of a formally explicit definition. The same sort of procedure is followed among scientific investigators. Observational reliability is sufficient if it does not lead, in practice, to dispute. Occasionally we encounter borderline cases, but most of the time we deal with observations that are

unambiguous enough so that no practical problem presents itself. In fact, dispute about the facts does occur sometimes in science, though rarely.

Explicit methods of establishing reliability of observers

Commonly we have sufficient experience with the class of observations dealt with in an experiment so that no explicit demonstration of interobserver agreement is necessary. If a rat is pressing a bar, and each pressing action causes a counter to increase its reading by one digit, there is little question that observers will agree on the counter reading. In other instances, however, it is necessary to give evidence of the reliability of observations. For example, Atkinson and McClelland (1948) placed groups of naval submarine-school trainees under varying degrees of food deprivation, and then recorded from the groups the frequency with which Thematic Apperception Test pictures elicited responses reflecting hunger. What is to be counted as a "response reflecting hunger"? Milking a cow? Bees lighting on flowers in a meadow? In order to be sure that a "response reflecting hunger" constituted an *observable event* in the scientific sense, it was necessary to arrange for more than one judge to observe the same phenomenon *independently*. Then the degree to which these judges agreed on the presence or absence of the event of interest was measured. Frequently the degree of agreement between observers is established by a statistical procedure. We might calculate a correlation coefficient between judgments. But it is probably more common to state simply the percentage of judgments on which independent observers agree.

Observation and measurement: The level of scaling

Measurement is a key activity of science. There are several different levels of measurement. Scales of measurement are usually classified into four types: *nominal, ordinal, equal-interval,* and *ratio.*

A *nominal scale* is simply one where a given individual, thing, or well-defined set of things can be placed reliably in one and only one category. Much psychological research is done with the independent variable scaled only nominally. For example, a study that compares the effects of liquid and solid food (see Chapter 13) is dealing with a nominally scaled variable. Liquid foods used in the study may be identified reliably. So may the solid foods. Each may be placed reliably in its appropriate category. Neither of the two types of food ranks above the other. An experimenter studying the effects of shock- versus food-motivation is similarly using a nominal scale.

The nominal scale meets the minimal scientific requirement of inter- and intraobserver reliability. We speak of it as a scale of measurement, but it is the lowest level of such scales. It is better to have scales at a higher level, since quantitative variations within the nominal catego-

ries may cause radical differences in results. For example, it would be appropriate to place roast beef sandwiches, cake, candy, or fish in the category "solid food." The range of liquid foods would be similarly broad. Depending on which pair of items we used in the two scaled categories, we might get quite different experimental outcomes.

The *ordinal scale* requires not only that we place items in a well-defined category, but also that the categories be capable of being placed in sequence. For example, we might be interested in studying the effects of birth order on some later performance. With respect to age the oldest child is above younger children, so the scale is ordinal. Suppose we studied the relationship between social dominance and levels of testosterone (male sex hormone) in the blood of women. If they rank in dominance 1, 2, and 3 . . . from most to least, the scale is ordinal. Notice that, although an ordinal scale stipulates the order of scaled items, it does not necessarily stipulate how far apart they are. For example, the most dominant woman might be much further above the second most dominant than is the second most dominant above the third.

With *equal-interval scales* we stipulate an arbitrary zero point and a unit of measurement, which remains constant over the entire scale.

Key Ideas Box 1.2: The nature of science

Philosophers of science have made numerous characterizations of the scientific method or methods. Much of their work has been based on an analysis of well-developed sciences such as physics. We will emphasize only a few central criteria. For psychologists, the most important method of science is the use of *measurement.* Scientific psychology was born when psychological events began to be *measured.*

The term *measurement* includes not only the high-level measurements found in more fully developed sciences. Mere reliable identification of an event is a form of measurement. The scientist must identify events consistently from one time to another and must do so consistently with other observers. It is in this sense that inter- (between) and intra- (within) observer reliability is the backbone of science. Facts must be reliably measurable. *Theories* must be reliably testable. A theory must be capable of being shown wrong in such a way that there would be no argument about it. It must permit *risky predictions.* A theory must enable one to say "Of all conceivable outcomes, these particular ones will occur." If it fails to do so, it is of no value, since it fails to help in predicting events.

Thus, when we mark off equal numbers of units, we mark off equal intervals. The Celsius and Fahrenheit scales of temperature are often cited as typical of equal-interval scales. Each has an arbitrary zero point, though not the same point. A unit is defined, but the Celsius unit is 1.8 times the Fahrenheit unit. Each can measure negative "below-zero" values.

The *ratio* scale has a zero point that is not arbitrary. Once you get down to zero grams or zero centimeters, that is all there is. It is not a matter of arbitrary convention. The units on a ratio scale are, of course, equal. A gram is a gram no matter where you are on the scale.

Science and experimental science

Some approaches to the study of nature deserve to be called "scientific" even though they do not entail the use of experimental procedures. Thus our subject matter is scientific psychology, but more precisely it is *experimental* psychology. How do the two differ? The answer is that experimental scientists manipulate and control the variables of interest, whereas nonexperimental scientific psychologists content themselves with observing and recording events as they occur naturally.

NATURALISTIC OBSERVATION

An example of a scientific discipline that is not usually experimental is astronomy. Astronomers rarely manipulate the factors involved in their study. Sometimes behavioral scientists also rely on observation and recording without experimental intervention. There may be considerable advantage in passive observation prior to experimental manipulation. Ethologists[1] commonly take great pains to describe the behavior of various species as it naturally occurs and regard this as essential to the development of a behavioral science. Sizeable books have been written on the life and habits of such species as the herring gull or the mountain gorilla. On the basis of these descriptions a great deal of intelligent inference can be made about why organisms behave as they do.

EXPERIMENTAL MANIPULATION

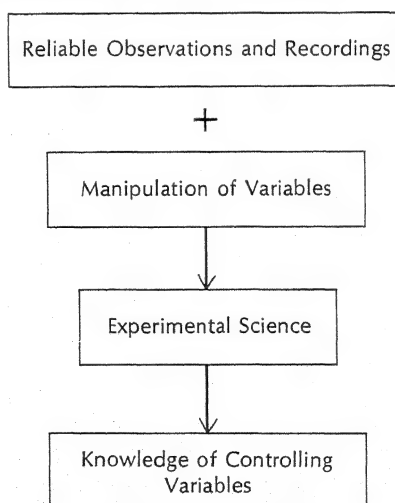
When scientists want to know which factors actually control a given behavior, they must introduce active experimental manipulation. They can decide that a given factor influences variations in behavior *only if*

[1]Ethology is a branch of biology in which animal behavior is studied, often in its natural setting. A delightfully entertaining introduction to it may be found in a book by the man who did most to create it—Konrad Lorenz (1952).

they discount the influence of all other factors. They discount the influence of these other factors by showing that variations in behavior still occur when the factors in question are *absent* or are *held constant.*

For example, ethologists have often asked which of the many features of an organism actually control the release of patterns of behavior such as courtship rituals or fighting. They have moved then from naturalistic observation to experimentation. To illustrate, an experimenter might find that for certain fish the presence of a red belly and of a certain orientation of the body is responsible for the initiation of fighting. How can this be determined? One way would be to construct artificial models that resemble an opponent in some respects but not in others, and then watch for the fighting pattern. If the pattern of the scales, movements of the gills, or shape of the eyes can be eliminated or changed without preventing this aggressive behavior, the experimenter concludes that these variables cannot be responsible for it. If, on the other hand, a feature such as the red belly precipitates fighting, even when other features differ markedly from those of the natural opponent, then it seems that "red belly" is an important controlling factor.

In general, it is by means of experimental manipulation that we are able to *specify exactly the conditions under which an event occurs.* The development of this method is summarized in Figure 1.3. When we restrict ourselves to natural observation, we cannot tell which of the many naturally occurring changes that take place at the same time actually controls the event of interest.

FIGURE 1.3 Experimental science combines reliable observation with manipulation of variables. It yields knowledge of controlling variables.

A PRELIMINARY STRATEGIC QUESTION: SHOULD ONE CHOOSE THE EXPERIMENTAL METHOD?

A primary strategic issue is whether to select the experimental approach to psychology. If one feels the objectives of experimental psychology are worth pursuing, the selection of experimental method is thereby determined, since it seems most unlikely that any other method could lead to these goals.

But some people fear the application of rigorous experimental methods to psychology. They feel that attainment of the objectives would lead to a sterile, controlled world shorn of the delights of spontaneity. Such fears are groundless, since even physics does not take away the unpredictability of a leaf falling to the ground. Prediction and control *in principle* imply only that a phenomenon can be reproduced when all the conditions of its occurrence have been specified and controlled. This is hardly applicable to specific situations in everyday life.

But should we try to control people at all? Why not let them be free? An option of this kind does not exist for us. There will be control. The only option is to decide whether it will be systematic or chaotic. Society controls behavior now. It does so in ways that are often clumsy, and that have undesirable side effects. For example, the methods of behavioral control most societies have evolved are largely punitive. Experimental science will let us know more constructive methods of control.

Further, the expansion of our understanding of how to control behavior will have unprecedented desirable consequences. There are

Key Ideas Box 1.3: Observational science versus experimental science

We can be scientific without being experimental. Many sciences rely heavily, or even entirely, on observation. They have necessary characteristics of science, observation and measurement. But they do not include active manipulation of the events studied. Astronomy was, for many centuries, purely observational. Such methods are valuable, but fail to discriminate between instances in which two things merely happen to occur together regularly and those in which there is a controlling relationship between the two events. In order to know which factors control observed events, we must turn to experimental manipulation.

many situations in which human suffering could be alleviated if we had a greater understanding of psychological control.

Finally, there are no methods of comparable power. People often think that common sense approaches to psychological helping are superior to the more rigorous approaches, but this is illusory.

Intuitive versus scientific procedures

THE MAIN ADVANTAGE OF THE SCIENTIFIC APPROACH

What are the advantages of a scientific approach? Put simply, the main advantage of a scientific attack on phenomena is that it makes *progress*, in particular *practical progress*, possible. Practical progress is unlikely without agreement on the facts because disagreement prevents integrated action. In science the rules according to which we can convince each other are relatively clear. Hence, to a great degree, we can reliably compel concerted action in controlling phenomena. For example, there are few people remaining who would propose a quack cure for polio. Science provided Dr. Salk not only with the means of discovering a preventive vaccine, but also with a method for convincing others that his technique would work.

RISKS ENTAILED IN NONSCIENTIFIC METHODS

Nature readily seduces us into believing that we understand her laws without the discipline of experimental science. The history of human inquiry is replete with instances of erroneous belief that yielded only to experimental attack.

Presumably, no one likes to believe what is false, but it is worth keeping in mind that *erroneous belief can also destroy human resources, human happiness, and even human life.* Did you know that it was long believed that a wound had to be cut in order to assure that vile "humours" would gather at the surface to be removed? This meant that a lucky individual who had a clean, uninfected wound would be treated with slices from an unclean knife until pus formed! The hapless patient would be *given* an infection.

It was not until modern times that a surgeon, John Hunter (1728–1793), noticed that survival rates seemed to be better in wounded soldiers deprived of treatment. And the surgeon did not content himself with what *seemed* to be. He ran a controlled experiment in which some people received the "treatment" and others did not, then compared their recoveries. It is doubtful that a mere impression could have served to break down such a longstanding belief, but an experimental demonstration could—and did.

Contemporaries tend to dismiss historical accounts of erroneous belief on the grounds that people used to be foolish, gullible, and prone to superstition. But truth comes to us no more easily today than it did in the past. Nonscientific psychology may be fascinating, but it is on ground no more firm than that of nonscientific disciplines in the past.

Dangers of intuition in psychology

Today, the treatment of mental disorder rests heavily on the individual talent of the one doing the treating—on "clinical intuition." But clinical intuition is so called precisely *because* it is unreliable, because one observer cannot agree with another. The intuitive approach is much more direct than a scientific approach and seems often more in tune with "common sense." But remember, common sense told our ancestors, who were no more stupid than we are, that it was silly to suppose that there were people underneath us on the earth, as the absurd notion that the earth was round would imply. Obviously they would fall off! The chief danger entailed by the intuitive approach is that it makes permanent self-deception possible. The situation of one who relies on intuition instead of science is similar to that of the fellow who thinks he is the greatest baritone in the world, even if no one appreciates his voice. There is no method of correcting such errors, since opportunities for external checking are missing.

If clinical intuition were enough without scientific controls, bloodletting would be a well-founded procedure, for centuries of clinical judgment support it. Many an astute barber-surgeon went through life convinced that bloodletting was an effective means of treating various diseases (even though at the time of the bubonic plague estimates of the amount of blood in the human body were so far exaggerated that people were literally bled to death). Of course the patients sometimes died, but they sometimes got better. No treatment is foolproof, after all! The trouble with the method of the bloodletting physicians was that they had no *objective criterion* of the effectiveness of their treatment—or at least they failed to put it to an objective test. They failed to make risky predictions and test them. Eysenck (1952) has pointed out that psychotherapists are doing exactly the same sort of thing today. They are convinced, on the basis of their clinical experience, that their therapy works, but they have not done the necessary experimental tests that would establish that they are correct.

A perspective on nonexperimental knowledge

My purpose in the previous sections has been to emphasize the value and importance of experimental method. Therefore, I stressed the advantages of experimental method and the disadvantages of other methods. Yet I, like everyone else, must often live and act on the basis

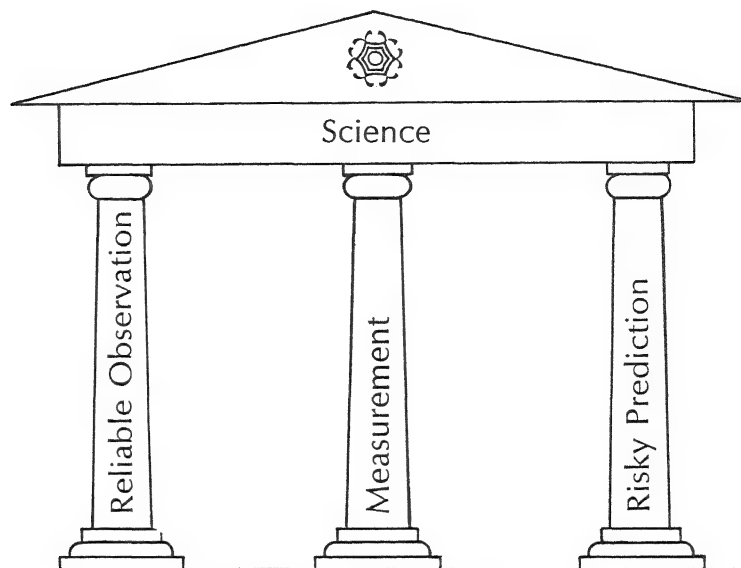



FIGURE 1.4 Foundations of science.

of much less reliable information than experiments provide. The ancient Greek notion that wisdom dictates using the best available knowledge in a given area continues to be a good rule of thumb.

Experimental method is, at bottom, an elaborate extension of reality-testing—checking out our ideas and fantasies to see if they hold up in the real world. It seems to be a superb elaboration of reality testing, though other methods might do very well under appropriate circumstances. For example, the methods of nonexperimental, observational science have been of very great service. And in everyday life we may have to settle for much less than that. But experimental methods can be extended much further than most of us realize. It is difficult, within our cultural traditions, to learn to examine every question for its experimental potential. Whenever possible, we should try to do that. And experiment is possible far more often than we realize. When it is not possible, let us accept the best knowledge available.

What about occult or esoteric knowledge? This would appear to be the polar opposite of knowledge based on experiments, yet reliance on such knowledge today is widespread. Even such forms of gaining truth can often be put to the test of science. Hypnotism is a good example of a phenomenon which made its way from the occult to the scientific. For some people, occult systems are merely sources of entertainment. But for those who take them seriously, reality testing is important. Why not experiment with these systems? If we can predict the degree of advancement of Zen disciples by measured brain waves (Kasamatsu &

Hirai, 1966), where are our limits? Attempts have been made to put astrology to scientific and even experimental test, with most interesting results (Gauquelin, 1973; West & Toonder, 1973).

In some cases, no public test of occult knowledge would appear to be possible. There is an occult tradition that "Knowledge is power and knowledge shared is power lost." The penalty for revelation of secrets among the ancient Pythagoreans was death. But even within these systems some methods of risky prediction can be identified. To what extent they are used we, as outsiders, cannot say. If they are not used effectively, the dangers of intuition are great.

The main disadvantage of experimental method is that it is slow and difficult. But the quality of the results will often make the effort worthwhile. Experimental checking will often show that even our best-founded ideas are mistaken. And we have not *begun* to extend the experimental method over its potential domain.

EXPERIMENTAL PSYCHOLOGY AND TRADITIONAL NOTIONS OF MIND AND BODY

The attitudes of experimental psychologists seem to me to run against traditional notions of mind and body. Actually, this may or may not be the case. I know of no adequate survey of the attitudes of experimental psychologists about such matters. Some psychologists take strong philosophical positions in favor of psychic determinism (that there is no free will) and materialistic monism (that all things, including mind, are material). But these are *philosophical,* not scientific, issues. Such questions cannot be answered within the framework of scientific method.

In fact, it is not hard to find experimental psychologists who are not determinists in the philosophical sense or who are dualists (mind and body are distinct substances) or even mentalistic monists (the universe is composed of only one substance, which is mind). Science has a way of conforming itself to changes of philosophy. In fact, one source of the success of scientific method has been its letting go of such philosophical questions.

I suspect that most experimental psychologists spend little time thinking about determinism or the mind-body problem. They do their experiments and leave such questions up to the philosophers. Scientific psychologists *do* adopt the working hypothesis that psychological events are lawful. This is the only working hypothesis that justifies doing scientific observations and experiments. But the possibility that there is ultimately a residuum of indeterminacy is of little concern within the framework of science. Physicists now accept indeterminacy with grace; so may psychologists.

Key Ideas Box 1.4: Advantages and disadvantages of scientific, experimental approach to psychology

Science uses methods that produce maximally reliable, repeatable findings. It is self-correcting. Since its concepts are regularly put to the test of observation, it makes the detection of errors most likely. In its experimental form, it tells us which variables actually *control* the occurrence of psychological events, as distinct from those variables that are only *correlated* with the events. Science makes progress, especially practical progress, possible. Since its methods are public, its discoveries are likely to be generally accepted and acted upon.

The main disadvantage of science in dealing with everyday life is that its methods may be slow and difficult to apply. Thus, we often must rely on other forms of knowledge. Yet experimental method can be extended much further than most of us realize.

Methods that lack the reliability and self-corrective measures of science make us highly subject to error. Worse, they may allow us to remain in error permanently. Nor do they often permit us, even if we are right, to convince others and get them to act.

The point is that, although any psychologist may have very strong opinions about philosophical matters, one does not speak as a scientist when one shifts to such questions. One may speak as a philosopher, or simply as a human being. But all of the familiar philosophical points of view can be reconciled with experimental psychology.

Controls, variables, and relationships

Essential to establishing the effectiveness of a maneuver is properly *controlled* experimentation. If we are to establish the efficacy of a given procedure such as psychoanalytic therapy, bloodletting, or adding a catalyst to chemical reagents, it is necessary that we introduce *controls,* which enable us to determine whether the observed events following it are actually due to our procedure, or whether some other factors are responsible for the events. We must make observations under two conditions: an experimental condition and a control condition. In the *experimental condition* the procedure whose effect is to be measured is introduced (for example, we arrange psychoanalytic therapy for a group of patients). Under the *control condition,* all factors are made identical to those under the experimental condition, but the experi-

mental treatment is omitted (for example, control subjects submit to all factors, such as coming to an office, except psychoanalysis).

As a second example of the use of controls, suppose an experimenter wanted to know whether the caffeine present in coffee improves reading comprehension. An uncontrolled procedure for answering this question would be to have people drink coffee and then take a reading comprehension test. But this would be a woefully inadequate method. The experimental treatment of introducing caffeine-containing coffee needs to be supplemented by a control condition. For example, some subjects (the experimental subjects) could be given regular coffee and other subjects (the control subjects) could be given a coffee with the caffeine removed. The reading comprehension scores of these two groups of subjects could then be compared. If enough subjects were used and they were assigned to the experimental and control groups without bias, a solid conclusion about the effects of caffeine in coffee on a particular test behavior of a particular population could be derived from comparison of their scores.

Any difference between the experimental and the control condition must be due to the introduction of the experimental treatment, if this is the only point on which the two conditions differ. Since the experimental and control subjects are often two different groups of individuals, we often speak of the *experimental group* and the *control group.* It is probably better to think in terms of experimental and control *conditions* or *treatments,* however, because we often use the same subjects "as their own controls." That is, we submit the same individuals to both the experimental and the control condition at different times and compare their reaction under the one condition to their reaction under the other. Under these conditions, there are experimental and control treatments, but there are no separate experimental and control groups. When we design an experiment this way, special precautions must be taken to evaluate the effects of the *order* in which we present the two conditions. It may make a considerable difference which treatment comes first. Methods of dealing with order effects will be discussed later.

CONFOUNDED VARIABLES

Although it is not always readily apparent, the reason why control conditions must be introduced in an experiment is to avoid what are termed *confounded variables.* A confounded variable is an uncontrolled factor—one that is allowed to vary along with the experimental treatment, making it impossible to tell whether any observed changes are due to the experimental treatment alone, or whether the confounding factor has made an important contribution.

For example, we might attempt to determine the effect of a drug on classroom performance by giving the drug to the women in a class and giving sugar pills to the men (the "fake drug" sugar pill is termed a *placebo* or, less frequently, a *nostrum*). Whatever differences we observed could be due to our experimental treatment, but they could also be due to differences between the sexes in classroom performances, with or without the drug. Sex is therefore a factor that is confounded with the experimental treatment. Note that we do not have to know for certain that sex has an effect on our observation in order to consider it a confounded variable. The mere possibility of its having such an effect is enough. *Any factor that is uncontrolled in an experiment and that is permitted to vary along with the experimental treatment is a confounded variable,* whether or not we know that it is actually making an important contribution.

INDEPENDENT VARIABLE, DEPENDENT VARIABLE, AND FUNCTIONAL RELATIONSHIPS

We want to eliminate the influence of any confounded variables because our goal is to discover the effects of our experimental treatment alone. For example, in Figure 1.5, the inadvertent pressure by the experimenter on the subject's foot should be eliminated so that it would not affect the response. We want to determine the effects of an *independent variable* on a *dependent variable.* When an independent variable has an effect on a dependent variable, the relationship between the two variables is called a *functional relationship.* The term "independent variable" corresponds roughly to the more familiar word "cause" and the term "dependent variable" corresponds approximately to the commonly used word "effect."

If we deprive a rat of water and measure the effect of this by the amount of water it consumes after the supply is restored, deprivation is our independent variable and amount of water consumed is the dependent variable. If we place an ant on a frying pan, gently warming it, and measure its rate of movement as a function of increase in temperature, the temperature changes constitute an independent variable and the rate of movement is a dependent variable. From these examples it is clear that the independent variable tends to be something that is manipulated or varied *directly* by the experimenter, whereas the dependent variable is manipulated only indirectly and as a result of its connection with (dependence on) the independent variable.

Why do we insist on introducing the new terms "independent variable," "dependent variable," and "functional relationship" when we already have the more familiar "cause," "effect," and "causal

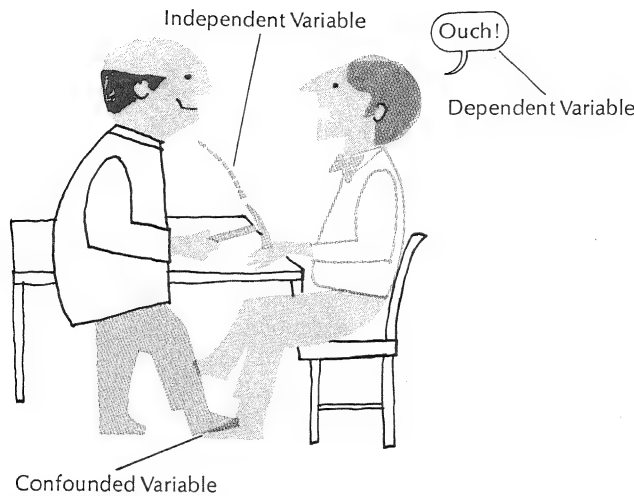



FIGURE 1.5 Relationship of independent, dependent, and confounded variables. The experimenter deliberately manipulates the independent variable and indirectly produces a change in the subject's behavior. If a confounded variable is inadvertently introduced, there is no way to tell whether the independent variable or the confounded variable produced the behavioral change.

relationship"? "Causal relationship" says more than we sometimes intend—it is a less general phrase than "functional relationship." Certain functional relationships are not causal. If we study grades in college as a function of performance on an IQ test, we find a functional relationship, but not necessarily a causal one. If the application of heat to the bottom of a frying pan causes an ant imprisoned therein to move faster, removal of the heat would presumably slow it down. But the achievement of a high score on an IQ test does not *cause* good academic performance. This would imply that one's not having taken the test would diminish one's academic performance. Presumably the relationship between IQ test performance and academic achievement is due to some third factor or set of factors. It is clear, then, that there is a difference between claiming to have established a causal relationship and maintaining that a functional relationship has been discovered.

Usefulness of terminology

We are now familiar with part of the fundamental vocabulary that is useful in scientific communication. The ability to use these terms uniformly is more important than it might seem. Not only does it help communication, but our ability to label independent, dependent, and confounded variables will help us identify these variables. As a consequence, the functional relationship under investigation will be easier to understand. Such verbal helpers can provide a sort of skeleton key to an otherwise bewilderingly complex experiment.

DELIBERATE CONFOUNDING

Independent variables in the real world are not simple and unitary. They have many facets. Since an independent variable is complex, we can see the course of scientific analysis as one of deliberate confounding followed by gradual elimination of confounded variables. An experimenter typically begins with a rather complex independent variable, and then becomes more and more analytic about the controlling factors as experimental sophistication increases. Even something as simple as the delivery of a food pellet actually involves simultaneous variations of many different dimensions. There are sensory aspects of a pellet; there are nutritive aspects; and there are consummatory response-inducing aspects, to name but a few. In the usual experiment, all factors are varied at once and we cannot say which one actually controls the behavior. Many experiments have been done to control one or the other aspect of food reinforcement. For example, experimenters have given nonnutritive, tasty substances like saccharin, thereby controlling for the effects of nutritive value (Sheffield & Roby, 1950). Others have tubed nutritive materials directly into the stomach, bypassing taste receptors (Hull, Livingston, Rouse, & Barker, 1951). These more analytic experiments illustrate the elimination of confounding of variables knowingly confounded by other experimenters.

One may question the validity of referring to the deliberate manipulation of complex independent variables as "deliberate confounding." This terminology is commonly used, but it is important to keep in mind that there is a big difference between (1) knowingly varying a complex of variables and then attributing any effects to some or all aspects of the complex and (2) the more traditional meaning of confounding. In the latter case, effects are uncritically attributed to one variable when another, unnoticed variable may have produced them. If we want to postpone analysis of the aspects of a food pellet that are rewarding to an animal, we make no mistake by acknowledging that it is not clear which aspect of the complex is responsible for the effects. It would be very different to argue that "nutritive value" was being manipulated and ignore the presence of other dimensions that vary. It is a matter of reserving judgment versus making an incorrect judgment. In any case, whether we refer to the procedure as deliberate confounding, as manipulation of complex variables, or by another name, it is a useful and commonly employed procedure.

Early versus late control

The mainstream of experimental psychology has been analytic, and this has led to an almost exclusive emphasis on a strategy of experimentation that involves cutting any problem into its smallest possible components before going to work on it. The attempt has usually been

made to control as many variables as possible, as early as possible, in the development of a new research area. When Ebbinghaus (1885) began his research into the experimental psychology of memory, he used "nonsense syllables" in order to minimize the influence of an important controlling variable—the number of associations already attached to verbal material. Psychologists have emphasized the use of animals and have restricted the use of human subjects to simple experimental situations—largely because of the tradition that control must be introduced as early as possible when a research area is being opened. Some people believe that this emphasis on control leads to

Key Ideas Box 1.5: Controls, variables, and relationships

In an experiment, we are looking for a relationship between two variables, the independent and the dependent variable. The *independent variable* is manipulated directly by the experimenter. The *dependent variable* in psychology is always behavior. It is manipulated indirectly, by the variations of the independent variable. Thus, if we control how many hours a group of people have been fasting, then measure their preference for sweet versus salty foods, the number of hours of fasting is the independent variable and the food preference is the dependent variable. If values of the dependent variable change as a function of changes in values of the independent variable, we say they have a *functional relationship*.

In order to isolate effects of an independent variable, we must have *controls*. A control condition ideally is one that differs from that under which the independent variable occurs in only one way—the independent variable is absent. Then any results that occur in the *experimental treatment condition* (the independent variable is present), but not under the *control treatment condition* (the independent variable is absent) may be attributed to the independent variable.

If something other than the independent variable is allowed to vary along with that variable, it is a *confounded variable*. Under such conditions, there is no way to tell whether observed behavior is due to the independent variable. Controls must be run to find out.

Sometimes we allow confounding on a temporary basis in order to see if a whole cluster of variables has an effect on the dependent measure. This is called *deliberate confounding*. Controls must then be run later to identify which are the variables responsible for observed effects *(controlling variables)*.

elimination of the most important features of complex human behavior.

In recent times there has been an increasing realization that confounding is always present in experiments, that deliberate confounding may even be a good experimental strategy under the right circumstances, and that a gradual process of its elimination is very common to the unfolding of an experimental area.

There is a lot to be said for dealing with complex variables first and then analyzing the function of their components as needed. If you can predict and control behavior without the analysis, that is all to the good. If you need further analysis for more exact prediction and control, you can introduce the necessary controls as required. Psychology would have developed very differently if the strategy of late rather than early introduction of analytic controls had been used. It is interesting to see what a different conception of memory was developed by Bartlett (1932), who used complex stories and visual materials to be memorized instead of nonsense syllables. He saw a directedness and orderliness in the forgetting process, which is not obvious in simpler situations. In addition, his procedure had the advantage of being truer to life than the more widely used analytic procedure. Ultimately, we want to be able to predict and control human behavior in situations that humans encounter in their natural environment. This goal might be reached more effectively by studying complex variables first and introducing controls as needed, rather than by trying to predict behavior in complex situations from data obtained piecemeal in simpler ones.

Explanation and pseudoexplanation

Once you have established a functional relationship of the "causal" type, you can use it as an explanatory tool. If you want to explain changes in the pressure of a gas, you can say that they occur because of prior changes in temperature (assuming that such changes and no others actually took place). Thus you establish a functional relationship between the temperature and the pressure of gas. If you want to explain an increase in the tendency for an experimental subject to drink water, you can appeal to an observation of, let us say, a prolonged period of water deprivation. Again you know that a functional relationship exists, and you use it to account for an observed fact or "phenomenon."

There is a highly prevalent counterfeit of such scientific explanation, called a *nominal explanation,* a *tautology,* a *circular explanation,* or (by Skinner, 1953) an *inner cause* explanation. A nominal explanation is a kind of verbal muddle in which the purported independent variable is a mere alias for the dependent variable. In reality, there have not been

two distinct observations (corresponding to the independent and dependent variables) at all. Only the dependent variable has been observed and then renamed, the second name being treated as an independent variable. To cite a comparative example from physics, examine the following question: Why do apples fall toward the earth? The usual answer is, "Because of gravity." But this simplistic answer is a misinterpretation of the concept of gravity. Gravity is merely a *name* for *the fact* that bodies fall toward earth. What observation leads us to conclude that gravity exists? The observation that bodies fall. Hence, if we say that bodies fall because of gravity, we are really saying that bodies fall because bodies fall. Our independent variable is really an alias for the dependent variable we are trying to explain. This does not contradict the laws of physics. Isaac Newton, the author of the universal law of gravitation, was well aware of the distinction between this misuse of the concept of gravity and the proper use of functional explanation. This awareness was in fact one grave obstacle to his being understood by his contemporaries (see Andrade, 1958). Newton said: *Hypotheses non fingo* (I make no hypotheses). He used the word "hypothesis" in its earlier meaning. Today we use the term to refer to an explanation that we accept provisionally in order to put it to experimental test. Newton had in mind hypothetical entities such as "powers" and "forces," which were used as all too easy pseudoexplanations of events—the very type of explanation we are objecting to here. Newton used the terms "power" and "force," but he defined them in terms of *observable events*. For him, they were mere shorthand forms for functional relationships. In the words of a contemporary physicist:

> What is gravity?
>
> But is this such a simple law? What about the machinery of it? All we have done is to describe *how* the earth moves around the sun, but we have not said what *makes it go*. Newton made no hypotheses about this; he was satisfied to find *what* it did without getting into the machinery of it. *No one has since given any machinery.* It is characteristic of the physical laws that they have this abstract character. The law of conservation of energy is a theorem concerning quantities that have to be calculated and added together, with no mention of the machinery, and likewise the great laws of mechanics are quantitative mathematical laws for which no machinery is available. Why can we use mathematics to describe nature without a mechanism behind it? No one knows. We have to keep going because we find out more that way. (Feynman, Leighton, & Sands, 1963.)

In the behavioral sciences we are especially prone to slip into nominal explanations. Explanation in terms of what Ryle (1949) has colorfully called "the ghost in the machine" has been particularly

common in psychology. If the ghost-in-the-machine account were adequate, there would be little need for the laborious development of psychology as a science, since the development of a complete account of human action is quite simple within such a framework. For every observed action, we would merely propose an inner "power" of the mind as its explanation. We would always have at hand a ready explanation for any human action. But if we were to use this kind of "explanation," it would only make the phenomena to be explained less accessible, and the task of developing a functional explanation would still remain before us.

Two examples of the nominal explanation of human actions are discussed below.

THE PRIMITIVE ARTISTIC IMPULSE. Begin with the striking observation that even the most "primitive" humans spent a great deal of time on art forms, even though the time could be alternatively spent in the realistic struggle for food and other basic necessities. Observations made on primates in their natural habitats, up to and including the anthropoid apes, give little indication of any noteworthy artistic strivings, but even in the most hostile of physical environments humans spend long hours on the design of intricate patterns for their shields, their huts, and the like. How do we explain this tendency? If we happen to be addicted to pseudoexplanation, we might say, "It is due to a primitive artistic impulse." But how do we know that the proposed primitive artistic impulse exists? We "know" that it exists because "primitive" men expend their energies on artistry—because of the very observation that we want to explain. Again, the observation corresponding to the independent variable is identical to the observation corresponding to the dependent variable. Put symbolically, this account is of the form "Humans do A because they do A."

THE PARANOID PERSONALITY. George is suspicious, stubborn, and has delusions of persecution. Why? Because he is paranoid. Once again, "paranoid" is a mere name for his symptoms.

Avoiding pseudoexplanation

How can we avoid slipping into nominal explanation? It is not easy. We have to keep in mind that the minimum requirement for an adequate functional explanation is that at least two *independent* observations be made. There must be one observation for the dependent variable and another, *distinct* observation for the independent variable.

It is often hard to tell in a brief encounter whether a nominal explanation is being foisted on you. Sometimes an individual who uses rather suspect language will, if pressed, be able to provide the listener

Key Ideas Box 1.6: Explanation and pseudoexplanation

A functional relationship, once established, can be used to explain events. If a child learns quickly under a regimen of reward, we can say that this is because of a familiar functional relationship, well known in operant conditioning. However, we must have at least two separate measures in order to have a functional explanation. We must have a measure for the independent variable and a measure for the dependent variable.

People often look at a behavior and, with no further observations, attribute it to some unobserved independent variable. For example, some people will say that mothers protect their offspring because of "maternal instinct." But if we ask "How do you know they have a maternal instinct?" the answer is likely to be "Because they protect their young." The explanation is circular. In order to correct it, we need to observe the "maternal instinct" as something independent of the behavior to be explained. Then we might have a true functional explanation.

Circular explanations are often given for psychological events. One of the most fertile sources of such explanations is the invocation of unobserved inner properties of the mind. But these are in fact no explanation at all, unless measured.

with the necessary definition in terms of observable events. Psychologists who have suffered through unpleasant encounters with commonly used nominal explanations are apt to be frightened away by certain linguistic hobgoblins without bothering to determine whether the seemingly loose language is being used in a rigorous fashion. It may be necessary to go to such extremes if you have trouble getting to the core of an explanation in order to see it for what it really is. On the other hand, if you judge explanations merely on the basis of the language in which they are framed you might be taken in by an even more deadly form of nominal explanation—the kind that has been carefully cloaked in scientific jargon. Nominal explanations are more widely encountered in psychology than we would like to admit. Even respectable explanatory devices such as "reinforcement," or "generalization," or "the central nervous system" can be used in unacceptable ways. The selection of words is not of great importance. The important thing is our ability to put observational backbone into our language. The speaker who uses everyday or phenomenological terminology may be more rigorous than many verbally less colorful colleagues. If the

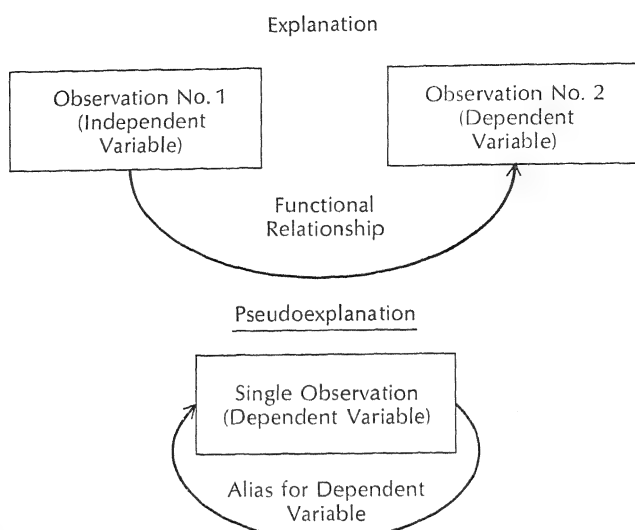



FIGURE 1.6 Contrast between explanation and pseudoexplanation. With explanation, there are at least two independent observations—one for the independent variable and one for the dependent variable. With pseudoexplanation, only the dependent variable is observed, and another name for the dependent variable is treated as though it were an independent variable.

minimal requirement of at least two distinct sets of observations has been met, the explanation is legitimate no matter how it is phrased. If that requirement has not been met, there is no real explanation. No matter how it is phrased, it will be a pseudoexplanation, as Figure 1.6 shows.

Study Questions

1. Define experimental psychology.
2. What does the text treat as the most important defining criterion (or criteria) of science?
3. What are the main methodological subdivisions of experimental psychology?
4. What procedures are ordinarily used to show that observations are scientifically adequate?
5. Distinguish between science and experimental science.
6. What is the main type of information that can only be obtained by experiment?
7. Why choose the experimental method?
8. What is the main advantage of scientific over intuitive procedures?
9. What is the main danger of relying on intuitive procedures?

10. What is a control? Give an example.
11. What is an experimental treatment?
12. What is an uncontrolled variable? Give an example.
13. What are confounded variables?
14. What is a placebo?
15. What is an independent variable?
16. What is a dependent variable?
17. What is a functional relationship?
18. What is the chief, immediate purpose of an experiment?
19. What is deliberate confounding?
20. What are the implications of using a research strategy involving late versus early introduction of controls?
21. What is a nominal (tautological, circular, or inner-cause) explanation?
22. Give an example, from your own experience, of pseudoexplanation.
23. What is the minimal condition for true functional explanation?

Exercises

1. Go to the library and obtain a recent issue of each of the following journals, in which psychological questions are answered experimentally. Page through them, noting the range of topics studied. Read the abstracts of those articles that have topics of interest to you. What is your impression of the scope of experimental psychology today? Is it broader than you thought? Could it be still broader?

   This exercise will help familiarize you with the psychological journals in your library, and will let you find out how to get them. It will also give you a view of experimental psychology unfiltered by your teacher or textbook writer.

   Bulletin of the Psychonomic Society
   Journal of Comparative and Physiological Psychology
   Journal of Experimental Research in Personality
   Perception and Psychophysics
   Journal of Experimental Social Psychology
   Journal of Experimental Child Psychology
2. Take some time to think about what kinds of psychological questions you *personally* would like to see answered. Forget about school. You probably would not be taking a psychology course if you did not wonder about some psychological matters. Think of your own questions before the trained psychologists impose *their* questions on you. Make a list of these questions. Can you see clear ways of answering them through psychological experiments? Come back to your list later in the course. Maybe

new ways of answering them will occur to you. This exericse may help you to see the use of learning about experimental method in psychology.

3. Select one of your own behaviors and describe it as precisely as possible. The number of cigarettes you smoke each day, the number of times you sit down, the number of times you say "uh" are examples. Any clearly identifiable behavior will do. Having identified the behavior, estimate how frequently it occurs per day. Write down your estimate. Now record and graph it daily for seven days.

   How accurate was your estimate? Discuss the accuracy of your classmates' estimates with them. On the basis of your findings, what do you think of the need for measurement?

2 SAMPLING PROCEDURES AND THE GENERALITY OF DATA

Once, after a well-known psychologist had finished delivering a lecture on some very socially relevant research in human social psychology, a colleague of mine threw up his hand and snarled, "What's this got to do with the behavior of rats?!"

He was only joking. He is a long-time student of both animal and human behavior. But his one-liner illustrates a prevalent objection to the work of psychologists. Does it apply beyond the particular subjects used by the experimenter? This is one facet of the problem of the generality of data.

Types of scientific generality

There are many types of scientific generality—as many as there are facets of an experiment. Does a functional relationship determined with rats apply to pigeons and to people? Does the relationship at certain values of the independent variable still hold at other values? Will a relationship calculated from data obtained by using a Lashley jumping stand also hold for a Skinner box? Do principles induced from small group interaction observed in a laboratory also hold in a business organization? All such questions pertain to generality.

If a finding (functional relationship) holds in situations other than the one in which it was discovered, we say it has generality or can be *generalized.* The situations may differ with respect to subjects, apparatus, measures of the variables, or any of the many facets of the original experiment. Without generality, conclusions drawn from scientific data would be of little value.

If all functional relations held only under the exact conditions under which they were obtained, science would be in real trouble. Virtually every situation would be a "new ball game," and the ability to make predictions would be limited almost to the point of incapacity. Scientists could hardly be motivated to do experiments if they believed that functional relationships had such limited application. They are far more inclined to believe that observed functional relations will survive many modifications of the original experimental situation.

Given the obvious importance of scientific generality, it is odd to learn that researchers pay relatively little attention to data generality—the basis of valid extrapolation of findings. They are very conscientious about the reliability of their findings, but take little interest in identifying the generality of these findings. Such attitudes prevail largely because of the commonly accepted rationale for assuring the generality of data. That rationale is virtually impossible to reconcile with what scientists actually do, so they tend to ignore the problem entirely. The rationale I am talking about is the one that says experimenters must assure generality of data through sampling procedures.

POPULATION AND SAMPLE

A scientist who does an experiment wants to make statements about some target group of interest. This target group is called the *population.* If one wants to make statements about alcoholism, one's target population is "all alcoholics." If one wants to make statements about mental illness, one's population is "all the mentally ill." If one wants to make statements about learning in general, one's population is "all creatures capable of learning." You can see that experimenters usually could not bring the entire population into the laboratory. Instead, they must settle for a part of that population. The part of the population they actually study is called the *sample.* So the problem of generality of data boils down to a problem of making statements about a population on the basis of observations made on a sample from that population.

Avoid getting the impression that the problems of scientific generality apply only to the matter of extrapolating from a subset of experimental subjects to a larger set. However, such extrapolation is a particularly important and subtle issue in experiments with human subjects.

Representative sampling as a basis for data generality

Most experimental psychologists believe that the generality of data is founded in *representative sampling.* With representative sampling, experimenters make sure that the sample of subjects used in their experiment has essentially the same composition as the general population to which they hope to extrapolate. They make sure that the sample is *representative of* the general target population. So experimenters who wish to study the opinions of college students (and thus wish to generalize any relationships found to the college student population as a whole), should try to be sure that the sample of students they study is typical. If it is typical, then they can be confident of the generality of their data.

Suppose someone interested in sampling the opinions of college students were foolish enough to inquire only at a campus lecture on meditation or only at a right-wing political meeting. In each of these cases the interviewer could have an unrepresentative or *biased sample*, and the study would likely run into trouble if its conclusions were generalized to college students as a whole.

Most of us are not likely to be so silly as the hypothetical experimenter just described, although more subtle sampling biases can find their way into an experiment. Experimenters doing brain lesion experiments might select their subjects by going to a rat colony and grabbing the first rat they can catch. This practice might lead to their having a sample biased in the direction of sluggish rats. Masters and Johnson (1966) had great trouble with sample representativeness in selecting subjects for their work on human sexual response. At first, they tried using prostitutes. Not only was this an unrepresentative sample, but there was definite evidence that the subjects' sexual physiology had certain peculiarities stemming from the practice of their profession. Masters and Johnson then sought volunteers. Here, again, there is trouble with respect to sample representativeness. The investigators were well aware of this difficulty, and discuss it in some detail (Masters and Johnson, 1966, pp. 9–23).

A particularly striking case of overgeneralization from a biased sample was described by Hoffman (1968) in his book on male homosexuality. In considering whether homosexuality is a mental illness, several psychiatrists decided that homosexuals are indeed mentally ill. The psychiatrists based this judgment on the fact that all the homosexuals they encountered were mentally disturbed. But, as Hoffman points out, their experience with treatment of homosexuals was restricted to those who sought psychiatric aid. Presumably, all their heterosexual patients were also disturbed! Hooker (1957) did a study in which she obtained 30 homosexuals who seemed well adjusted. She matched them with 30 heterosexuals for age, education, and IQ, and

submitted all 60 subjects to a battery of psychological tests. Several of her colleagues were then asked to evaluate the test results, without knowing the sexual preferences. They were unable to distinguish homosexuals from heterosexuals. Thus, the generalization that all homosexuals are mentally disturbed was due to sampling bias.

Instances of overgeneralization from biased samples are easy to find in popular science articles, and even in many technical ones.

Methods of sampling

What do researchers do about the problem of sample representativeness? They try to see to it that each member of the population of interest has an equal opportunity of being selected or that the probability of being selected is proportional to the real representation of subjects of that sort in the population. With *random sampling with replacement,* each subject in the population has an *equal* and *independent* chance of being selected. One has as good a chance as anyone else, and one's selection at a given time has no influence on future selections, for one is as likely as any other subject to be selected for future samplings.

RANDOM SAMPLING WITHOUT REPLACEMENT. Sometimes it would be a mistake to sample with replacement. A subject, once used, may be changed in some way so as to become ill suited for future use. For example, an experiment might include a test that would be easier if taken a second time. Or there might be some gimmick in the experiment that is exposed to participants who are then irrevocably knowledgeable about it. In animal research there may even be permanent bodily changes, such as modifications of the brain.

Thus it is common to sample without replacement. Does this mean that the sample must necessarily be unrepresentative? No. The experimenter can make sure that the fundamental *sampling unit* is randomly selected. The sampling unit can be an individual subject, but it can also be a group of subjects. If you consider all the possible samples of a certain size that could be taken from the population, you can treat "samples of size n" as the sampling unit. Then the sampling unit can be selected randomly. In *random sampling without replacement,* each sample of size n has an equal and independent chance of being selected.

THE USE OF "RATIONAL PROCEDURES" IN SAMPLING. Some experimenters do not rely entirely on the chancy process of randomization in making sure that they get a representative sample. They use their knowledge of the population and simply see to it that the sample has the correct composition. Opinion pollsters commonly do that sort of

thing. They usually make sure that the various regions of the country, which they know to be important determinants of political opinion, are represented in their sample in a proportion like that found in the population as a whole. They often do the same thing with age, race, and declared political party affiliation. You should realize, though, that within these imposed limits randomization is generally used. These samples are merely placing what are called *constraints* on randomness.

If you rely on your knowledge of the population and take control of the sampling to assure that it is representative, rather than relying entirely on randomization, you are using "rational procedures" in sample selection. The rational procedure I just described is called *stratified sampling.* With stratified sampling, care is taken to assure that the proportions of certain key elements in the sample are the same as those in the population. For example, in selecting human subjects for an experiment, the experimenter might see to it that the proportion of women selected is the same as that found in the population as a whole, or might attempt to assure that whites and blacks are represented in accord with their proportional representation in the population. Selection is otherwise random. We must, of course, have certain information about the population in order to do this.

The details of sampling procedure vary, but their purpose is the same. They are designed to assure that simple random sampling of the sampling unit occurs. If this can be accomplished, representativeness of the sample will have been assured. In this way, the experimenter can have confidence that findings about it are general across subjects.

It might seem to you that rational procedures are better than randomization procedures. People are usually inclined to be uncomfortable about leaving things to chance when they can have control over them. But keep in mind that rational procedures depend on the accuracy of knowledge about the population. If you make a mistake about this, you may fail to get a representative sample.

PROBLEMS IN THE PRACTICAL USE OF SAMPLING THEORY. If experimenters use the sampling procedures described, data generality across subjects is assured, but it is hard in practical situations to adhere to the principles of sampling. If you want to study human memory, a random sample of the population to which you wish to generalize will require that every human being be included in the potential sampling population. Clearly, you would be hard-pressed to fulfill the conditions of the sampling procedure under such conditions.

In practice, experimenters typically use a kind of "catch as catch can" sampling procedure. A human learning study is done on a selection of volunteers from among students enrolled at a given university, rats are selected from a colony of albino rats available in the

experimenter's laboratory, or other convenient sources are used. The conditions required by sampling theory to assure representativeness are simply not feasible in the typical psychology experiment.

Another problem in the use of sampling theory as justification of data generality is that it applies only to generality across *subjects.* There are many other aspects of generality. How can an experimenter apply sampling principles across experimental designs, types of apparatus, and the like? Little attention is given to problems of this type, yet they are problems as important as those involved in subject sampling.

Despite the failure of sampling theory to justify most of what is done by experimenters, there is a tendency to adhere to it as a rationale for statements of generality. For example, a scientist may object to the use of rats because they are a species unrepresentative of the population to which we wish to generalize, yet fail to apply the same arguments to other features of the experimental setting.

Clearly, some other rationale for the affirmation of data generality is needed. An experimenter who can accomplish true random sampling certainly should do so. But if, as is commonly the case, this cannot be done, a different view of generality is required.

GENERALITY OF DATA AS AN EMPIRICAL QUESTION

The best way, even better than representative sampling, to find out whether data are general is to demonstrate by experiment. If you run an experiment with one pigeon as the subject, and someone suggests that the finding might be unique to that particular pigeon, the best answer is to run another pigeon. If you run an experiment on social behavior in a small group in the laboratory, discover regular behavioral patterns, and wonder whether similar patterns of behavior occur in comparably structured small groups in factories, the best approach is to repeat the experiment in a factory. If you have evaluated prejudice by means of a questionnaire and wonder whether different measures such as heart rate, pupillary dilation, or behavior when offered the opportunity for an actual act of discrimination will yield the same results, nothing could be more convincing than actually doing experiments to find out whether generality exists.[1]

Unfortunately the empirical approach to data generality has its limitations, just as the sampling approach does. If we are limited in our statements of generality to conditions identical with those situations we actually observed, our science will be virtually useless. Even if you show that a finding obtained in your laboratory can also be obtained in

[1]In all of these instances you would be using a procedure known as *systematic replication.* I have postponed a fuller discussion of replication until Chapter 4. Logically it could occur under "reliability" or under "generality," since, at a deep level of analysis, reliability is a special case of generality. But we need not go that deep here.

a factory, what about other laboratories and other factories? To make truly general statements based solely on actual observations, it would be necessary to experiment in every conceivable setting, using every conceivable method. Even if this could be accomplished, generality with respect to future situations would have to be based on something other than direct observation.

THE COMPLEXITY OF BASES FOR AFFIRMING GENERALITY

Judgments of generality are founded on a variety of considerations. When it is argued that future pigeons may not be like past pigeons, most experimenters react to the argument as a parent might react to a tedious child. Their tendency is to treat such objections lightly and to go on insisting that generality is an empirical question. They overlook the fact that there is a reason for their unwillingness to take seriously the possibility that the laws of nature will change today. In fact, there is a whole assumptional system within which scientists work, and one of the assumptions seems to be that nature displays some degree of stability.

Another point that is seldom mentioned is that the assumption of generality, in the absence of evidence to the contrary, is simply a *good strategy.* This is essentially what the *principle of parsimony* is about. The principle of parsimony states that scientists should not add new explanatory principles unless the evidence requires them. Thus, the simplest possible explanation, the one with the fewest parts, is preferred. At the very heart of science is the urge to reduce the complex world to a few basic principles.

Since scientists want to keep explanatory principles to a minimum, they are inclined to expect that a principle that applies in one situation will also apply in similar situations. Thus they tend to generalize first and ask questions later. There is an apocryphal story that the physicist Isaac Newton, who had been sent home from school to his farm during the spread of bubonic plague over Europe, was sitting under an apple tree and was struck on the head by an apple. He immediately had an "Aha" experience, and went on to discover the universal law of gravitation. He *observed* an apple falling on his head, but he intuitively concluded that there must be a certain law of attraction between all physical bodies on earth or in the heavens. He assumed that what occurred in one small place applied everyplace.

A similar parsimonious leap occurred when it was first discovered that a magnetic field is generated whenever electric current flows. There is always a magnetic field perpendicular to the direction of flow of any electric current. When the French physicist Ampere (after whom the unit of electrical current "ampere" is named) learned of this, he

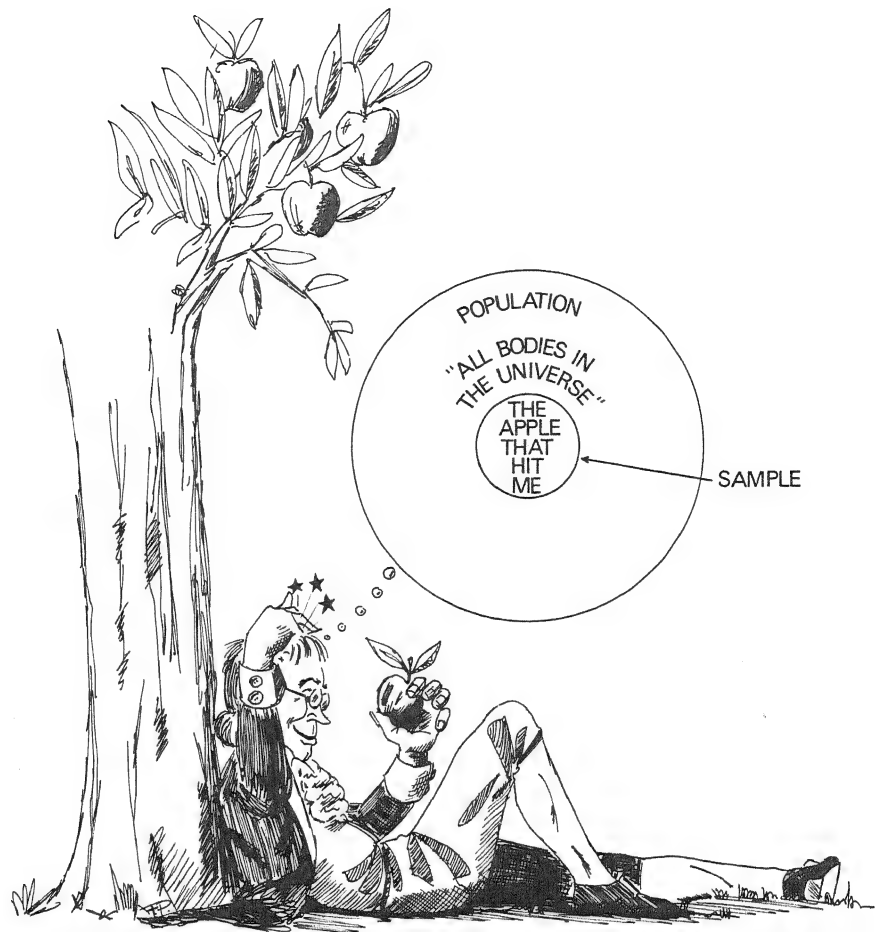



FIGURE 2.1 The relationship of Newton's population and sample. Should he generalize?

immediately supposed that *all* magnetism, even that found in permanent magnets, was due to the flow of electric current. This went far beyond the observations, since no one had observed current flow in permanent magnets. It was in accord with the principle of parsimony in that it assumed that only one set of principles explained both the properties of permanent magnets and those of electromagnets. And it proved to be a correct supposition, for we now know that there is indeed a net movement of electrons in those materials that are permanently magnetized.

We generalize unless given evidence to the contrary. For example, given a single observation from a population, the best guess about the character of the remaining observations is that they will be the same as the first one. A simple example will illustrate the point. Suppose you

are asked to reach into a box and draw out an item. The item withdrawn is a black sock. If you are now asked, in the absence of any other evidence, to guess the nature of the next item you will draw, what will you say? Many people will simply say, "I haven't got the foggiest notion," but that strategy is not a good one. It would be better to guess and say, "a black sock." Why? Because the items most frequently represented in the box are the ones that were most likely to be drawn in your one attempt. Thus, by guessing that future items will be the same as the one you drew, you can do better than by failing to guess at all, or by guessing "at random." Similarly, it seems preferable to assume that observations have generality than to assume that they do not or that there is no way of guessing.

Notice that the argument just given is couched in terms of *guessing.* This sort of thing will certainly not make anyone certain that a finding has generality. It simply means that, lacking anything better, generality is a good guess. Verifications that the guess is correct will strengthen our conviction of generality and falsifications will weaken it. Empirical evidence is always the most compelling source of information. To assume that guesses, even strategically good guesses, are absolute truths is to be guilty of hasty generalization—an error that has often caused scientists a great deal of trouble.

The special problem of experiments designed to affirm null statements

Experimenters sometimes design their experiments so that the main conclusion, if the experiment is successful, will be that an independent variable does *not* have an effect rather than that it *has* an effect. It is generally wise to avoid experimental designs of this type. This is not to say that it is entirely impossible to conclude with reasonable certainty that a variable does not have an effect. It is simply difficult. And sometimes it is easy to modify an experiment of this sort so that the conclusion can be that an independent variable *has* an effect. So why not avoid troubles and change the experimental design?

SPECIFIC OBJECTIONS TO EXPERIMENTS DESIGNED TO AFFIRM NULL STATEMENTS

There are many objections against designing experiments to affirm null statements. Some of them are too complicated to deal with here, but I will list the major ones. Null statements are often overgeneralizations.

Probably the most striking case against them is that experimenters sometimes make very slight changes in an experimental procedure and find that an effect is produced where none appeared before. For example, the great physicist Hans Christian Oersted regularly gave

Key Ideas Box 2.1: The generality of data

The generality of an experimentally determined relationship is its applicability to situations other than the one in which its supporting data were obtained, for example, with different subjects, in a somewhat different setting, with a new experimenter, and the like. An experimenter is interested in drawing conclusions about a target population (such as "all human beings") on the basis of observations made on a sample (such as "ten college freshmen"). However, generality must apply not just across subjects, but across all aspects of an experiment.

BASES FOR GENERALITY

There are various ways of establishing generality.

1. *Representative Sampling:* If the sample used in an experiment is representative of the population of interest, generality of data across subjects will be assured.

   There are many procedures to assure representativeness of samples. These include:

   Random sampling with replacement: each subject from the population has an equal chance of being selected and the selection of a given subject has no influence on *any* other selections. This means that a given subject may be selected more than once.
   Random sampling without replacement: each sample group has an equal chance of being selected from the population, and no subject can be selected more than once. This method is used when it would be unwise to employ a given subject more than once.
   Random sampling limited by rational procedures: the experimenter uses what is known about the composition of the population and deliberately arranges for the sample to have fundamentally the same composition.

   Limitation on representative sampling: It is almost never possible to obtain truly representative samples from the populations in which experimental psychologists are interested.
2. *Simple Enumeration:* With this procedure generality of relationships is determined empirically. If someone asks whether a finding applies in a new situation, the experimenter answers by trying the experiment out in the new situation.
   Limitation: Scientific findings are useless if they cannot be used to anticipate events not yet observed, for example, future events. Simple enumeration cannot enable us to generalize to situations not yet observed.
3. *Good Strategy:* The assumption of generality has been a basic methodological tool of science. It is identical with the generally accepted scientific *principle of parsimony,* which states that the explanation requiring the smallest number of principles (the *simplest* explanation) is to be preferred. When we assume generality, we assume that principles established in one setting also apply in other settings; thus we keep the number of principles to a minimum.

demonstrations showing that there was no observable interaction between electrical currents and magnetism. He would place a compass next to a wire, run current through the wire, and show that the compass needle did not deflect. He made the reasonable assumption that any magnetic effects would be in the same direction as the flow of electrical current. One day, after one of the demonstrations, someone thoughtlessly changed the orientation of the compass and threw the switch. The needle deflected! We know now that the magnetic field is not parallel, but at right angles to the flow of electrical current. A seemingly slight procedural variation changed the null finding to a positive one that launched the new field of electromagnetism.

In psychology, a classic dispute between Paul Fields and Norman Munn, which took place during the 1920s, was centered on affirmation of a null statement. It was widely accepted at that time that rats could not discriminate visual patterns. To say that they could not discriminate such patterns is to maintain that the independent variable (pattern) does not, under any circumstances, exert control over the organism's behavior. Fields (1931) devised a technique that enabled him to obtain pattern discrimination, but only after hundreds of trials. A diagram of his apparatus is shown in Figure 2.2. Munn (1929, 1930) attributed the discrimination behavior obtained by Fields to confounded variables. Several articles were written in dispute of this issue. Many experimenters had been unable to get rats to discriminate pattern, and were so convinced that their null findings represented the real state of affairs that they could not bring themselves to believe Fields.

Later, Karl Lashley (1930) invented a discrimination apparatus called the *Lashley jumping stand,* in which rats discriminated all sorts of complex patterns in very short order, commonly fewer than 100 trials. The Lashley jumping stand and some patterns discriminated by rats in it are shown in Figures 2.3 and 2.4. Lashley actually showed that rats could discriminate patterns in his apparatus when only a few

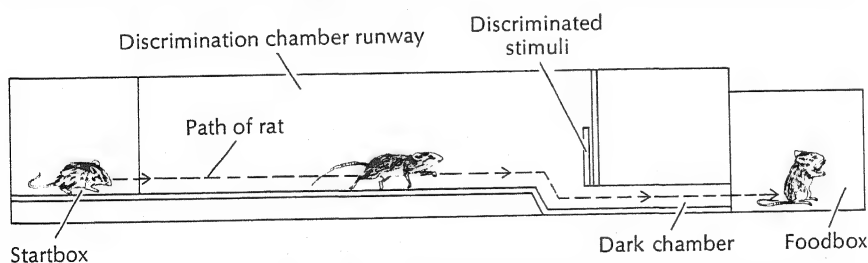



FIGURE 2.2 Schematic diagram of discrimination box used by Fields (1931) to train rats to discriminate visual pattern. The rat begins in the startbox, traverses the runway, at the end of which the stimuli that he must discriminate are displayed. He passes beneath the stimuli into a dark chamber, through which he reaches a box where there is a reward. (Adapted from Fields, 1931.)

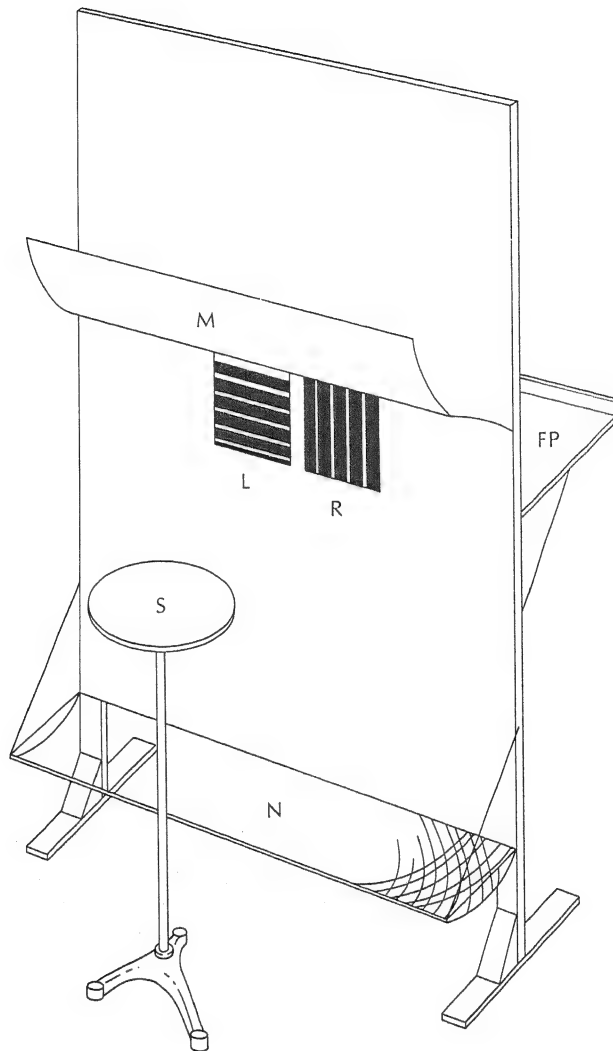



FIGURE 2.3 Apparatus used for testing discrimination of visual patterns. *S* is a start platform from which the rat jumps. *M* is a deflector for rats who jump too high. *L* and *R* are the left and right openings through which the rat gains access to *FP*, the food platform. The openings are covered by cards bearing the visual cues to be discriminated. The correct opening is covered in such a way as to allow the card to fall and give access to *FP*. The incorrect opening is locked so that a rat jumping to it will fall into *N*, a net. The position of the correct cue (right or left) is randomized so that the response cannot be made on the basis of a position preference. (From Lashley, 1930.)

hundred of their visual brain cells had been spared following brain surgery. Today, psychologists seem to have forgotten that there was ever any difficulty in getting rats to discriminate pattern—

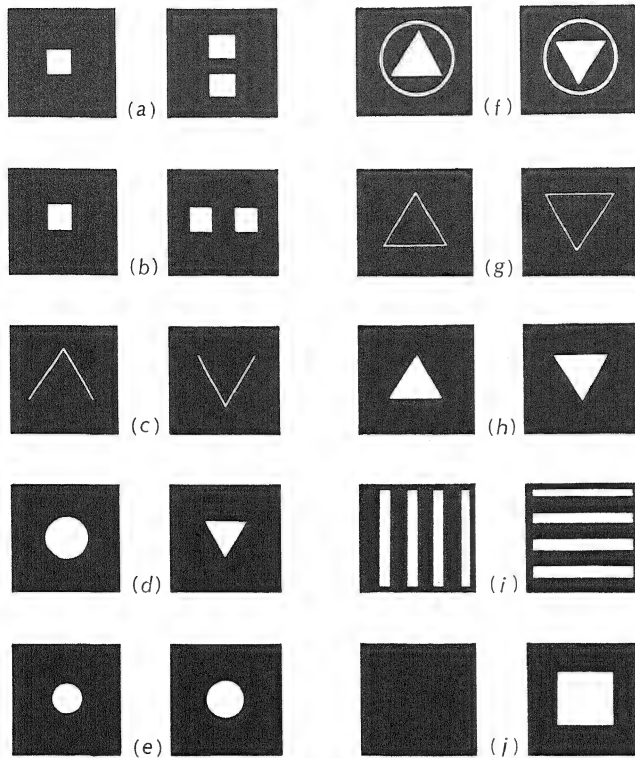



FIGURE 2.4 Pairs of patterns between which rats have been trained to discriminate in the apparatus shown in Figure 2.3. (From Lashley, 1930.)

contemporary psychologists accomplish this every day, using various types of apparatus.

STATISTICAL OBJECTION TO EXPERIMENTS AFFIRMING NULL STATEMENTS. Experimental psychologists know that they are sometimes wrong. Indeed, they go to great lengths to quantify the probability of their being wrong. The major tool they use to determine that probability is statistics. Thus, it is at the very heart of the statistically-oriented experimental psychologists' procedure that they be able to put a number on the probability of their committing an error. Unfortunately, this is not practical if they affirm a null statement. Later we will discuss the ways in which we identify errors in psychological experimentation (Chapter 4), and how probabilities are assigned to them. For now, it is enough to say that the probability of being wrong in deciding that an independent variable had no effect is called *beta,* and it is not a single number. It is a function, a whole curve. The information necessary to know which of the many points on that curve represents our particular

Key Ideas Box 2.2: Experiments designed to affirm null statements

An experimenter sometimes designs his experiment so that the main conclusion is likely to be that some experimental treatment had no effect. We call such a conclusion the acceptance of a null statement. Experimenters should try to avoid accepting null statements, since this is unduly risky.

ARGUMENTS AGAINST ACCEPTANCE OF NULL STATEMENTS

1. Scientists like to stipulate the probability that their statements are wrong. When experimenters reject a null statement, they can specify that the probability of their being wrong is some number, which statisticians have given the general name "alpha." When experimenters accept the null statement, they cannot specify the likelihood of error nearly as clearly, since the probability of this being wrong is not a simple number, but a function of several variables.
2. When experimenters accept a null statement, they usually make a stronger statement of generality than when they reject such a statement. An experimenter who says that a certain treatment has a specified effect usually knows full well that the effect will wash out under some conditions, known as the *limiting conditions* of the functional relationship. For example, a given teaching technique might be effective under normal conditions of room temperature, but at 140 degrees Farenheit—in a sauna—it probably would not help. When experimenters say that such-and-such treatment has no effect, they usually mean that it won't work under *any* circumstances. For example, people used to say that apes can never learn to talk. They meant that no set of treatments would permit them to use words and syntax creatively. Recent research makes this more than doubtful.
3. Historically, scientists who have accepted null statements have often had to eat their words. One scientist used to give lecture demonstrations showing that there was no relationship between electricity and magnetism. Somebody made a 90° change in his demonstration gear, and a marked relationship between electricity and magnetism became obvious. The same scientist thereby founded the study of electromagnetism.
4. If we were to treat failures to show the effect of a variable with an amount of respect equal to that given successes, then the laws of chemistry and physics would virtually all be refuted. My guess is that many attempts to replicate basic chemical and physical experiments are failures, since most of them take place in high-school and college laboratory courses where the students lack certain important skills, and even basic interest. The main point here is that all kinds of oversights can obscure the effect of a powerful independent variable. Hence, failure to get an effect is often trivial.

probability of error is generally not available to experimenters. Hence a major goal of statistics is lost by accepting null statements.

Experimenter effects

As the old saying goes, the last thing a fish will discover is water. It seems that experimenters were almost as tardy in discovering that one very important factor limiting the generality of experimental findings is the experimenter's own psychology. People have long been aware of the possibility of *experimenter bias.* This term refers to the tendency of experimenters—usually inadvertently—to influence or select their results in such a way as to invalidate them. Many familiar aspects of experimental procedure were originally introduced to cope with the problem of experimenter bias. For example, we suspect, as Darwin did (Wilson, 1952), that experimenters tend to be especially forgetful about facts that do not fit well with their theory or hypothesis. Thus scientists are trained to keep careful records of all their data.

Some experimenters get so wrapped up in their results that they want to throw out data that fail to fit their expectations. Recently some students who were conducting research with me had findings that, by and large, supported our expectations until the last few days of the experiment. They wanted to ignore the late data, arguing that, since the subjects were college students, and since the disagreeable data came near the end of the week, and since (they conjectured) more exams come at the end of the week, only the earlier data were valid. This would be a very dangerous thing to do. It is a mistake so common that it has been given a name all its own: "the early returns effect." It means that experimenters tend to build up expectations on the basis of the early data, then to bias the later data in such a way as to make them fit in with the expectations.

The procedures of experimental design have been devised to avoid such obvious mistakes. However, experimenters may still influence their results without meaning to. One way that I have seen many times lies in the calculation of statistical significance. Few experimenters enjoy doing the long, drawn-out calculations involved in testing the statistical significance of data. And few people like checking calculations by doing them more than once. I have noticed that an experimenter who has done a statistical analysis, if it comes out significant, smiles and quits; but if it falls short of significance, the experimenter frowns and starts mumbling about a possible error of calculation. Experimenters are more likely to check the calculations on data that they dislike, and can thereby bias the outcomes of experiments. This is easily corrected by making it a policy to check all calculations, painful though it may be to look for errors in calculations that have yielded results that make you smile.

These problems of experimenter bias are relatively easy to handle, but there are other experimenter effects that sneak up on you even when you are very careful. One is hardly surprised to learn that an experimenter can read a desired result into data if measures are vague. Nor is it surprising to learn that scientists, being human, tend to forget items of information that do not support their pet theories. But it *is* surprising to find that the attitudes of experimenters can influence experimental outcomes when measurement and recording are done by machine. For example, in one experiment involving the training of rats in completely automated operant chambers, the experimenters were led to believe that they had a "stupid" or a "bright" rat. The performances of the rats were in accord with the artificially induced expectations of the experimenters (Rosenthal & Lawson, 1964). Apparently the manner in which experimenters dealt with the rats outside the experimental chamber was able to cause major variations in the animals' performances.

The term *experimenter effect* is broader than the term "experimenter bias." Bias is something in the behavior of the experimenter. But even nonbehavioral cues and expectations can have an effect on subjects. The experimenter's appearance, personality, sex, social status, and the like may be important.

There have been many, many demonstrations that various characteristics of experimenters influence their data (Kintz et al., 1965; Rosenthal, 1966), and, though certain limitations in these studies have been pointed out (Barber & Silver, 1968), they have been convincing enough for us to conclude that the experimenter's personage should be treated as an important variable in an experiment. Oddly enough, it is the experimenters in the "hard-nosed" branches of psychology who seem most likely to neglect considering the experimenter effect. "Tender-minded" psychologists, such as researchers in clinical psychology, are more likely to be aware of the importance of such influences, probably because it is very obvious that human judgments play an important role in their research.

There are several ways of minimizing, or at least of detecting, experimenter effects. For example, it is possible to have several experimenters in the same laboratory conduct a given research project so that the influence of a given experimenter's individual characteristics can be assessed at the end of the experiment. Indeed, it is common for several investigators to participate in the execution of a given research project, but it is relatively rare for the personal influence of the investigators to be assessed with the same rigor as are the influences of more obvious variables. A second technique is to conduct experiments *blind.* This means that the experimenter administering the treatment does not know whether the experimental factors (variables) of interest are present or absent, and the person keeping records of the variables

reveals them only after the experiment is completed and its outcome judged. An experiment with human subjects is called *double blind* if neither the experimenter nor the subjects know which treatment they are receiving at the time of the experiment. For example, an experiment on the behavioral effects of a drug might be done by coding the vials of drugs and placebos so that no one knows which is which during the experiment. This can only be determined after decoding.

None of these procedures is foolproof, but at least they can be helpful. In any case, it is important to realize that characteristics of the experimenter must be taken into account very carefully and regarded just as seriously as any other potentially influential factor. There is little point in meticulously taking into account the subject's background, environmental surrounds, and all other factors that plague psychological experiments, only to allow the experimenter factor to control the situation.

Key Ideas Box 2.3: The experimenter effect and experimenter bias

The *experimenter effect* is a general name given to all the various ways in which an experimenter may personally influence the experimental results. The best-known experimenter effect is *experimenter bias,* in which the experimenter behaves in a way that changes the yield of data. This could be deliberate, but more often the experimenter does this without being aware of it.

A great deal of research has been done by psychologists showing that there are experimenter effects beyond those of bias. Even without exerting bias behaviorally, an experimenter can have an individual influence on the results. Such things as the age, sex, and social class of the experimenter are influential.

Though far from foolproof, a number of special precautions may help to minimize experimenter effects. The use of *blind* procedures, especially *double blind* ones, may be useful. A procedure is double blind if neither the human subjects nor the experimenter know when the experimental treatment, rather than the control, is being given. This is done by coding the experimental treatments in some way so that the information may be retrieved at the end of the experiment.

It is also useful to have more than one experimenter and to find out directly whether the results generalize from one experimenter to others. In order to do this the data must be analyzed in such a way as to detect the influences of different experimenters.

At this point a word of solace is probably appropriate. Experimenters, especially beginning experimenters, often become discouraged by the difficulty of designing valid experiments. The realization that they might themselves unintentionally influence their results may seem almost too much to bear. But the problem of coping with experimenter effects is neither more nor less formidable than the problem of coping with any other type of extraneous effect. Science has progressed quite nicely despite the imperfections of the world and its fallible human experimenters. The ideal experiment has never been done, just as the ideal life has yet to be lived. Fortunately, science has its checks and balances to prevent errors from being perpetuated indefinitely. The problem of whether a given experimental outcome is due to the particular experimenter is primarily a question of whether the data sought have generality. (Can a finding of one scientist be obtained by another scientist? Is it *generally* found by all scientists?) Questions of generality face the scientist every day and there are ways of coping with them. Experimental science is difficult, but it is certainly far from impossible.

Devising tactics for generality, reliability, and economy

In order to devise a tactic, you need to be aware of tactics already used and of logistic opportunities available. These matters are discussed throughout this book. However, there are certain very general things to be said about selecting tactics. We cover these general matters here, because they relate closely to the problem of generality.

SELECTING A SETTING

Experimenting in the target setting

Selecting the setting of your experiment is a critical step. The first possibility is to do the experiment directly in the target setting of interest. Experiments are seldom done in natural settings, but this can be done to considerable advantage. Some advantages are that:

1. It maximizes generalizability to the target setting.
2. It may minimize the artificiality and the intrusiveness of experimental settings.
3. It permits the full range of natural behaviors to occur.

The main disadvantages are:

1. It may make the exercise of experimental control difficult.
2. It may be logistically, including ethically, difficult.

Keep in mind that there are important distinctions between a *naturalistic experiment* and an experiment in a natural setting. In a naturalistic experiment, the experimenter does not have control of the independent variable, but waits for natural changes to occur in the independent variable. In contrast, the experiment done in a natural setting involves *manipulation* of the independent variable by the experimenter. The latter arrangement permits more complete isolation of the controlling variables.

It is also important to recognize that an experimenter working in a natural setting must still select behaviors to be measured. It is not enough to give a general description of what occurs in that setting. Behaviors must be measured if we are to have a scientific experiment. Mere casual description like that done by nature lovers and people watchers is not enough.

Experimenting in a model setting

If the disadvantages of experimenting in the target setting outweigh the advantages, it will be necessary to find or devise a setting that is a good model of the target setting. I use the term *model setting* to refer to anything that, when manipulated, can be expected to produce effects that can be generalized to those found in the target setting itself. When an experimenter brings people into a laboratory and arranges for group pressure to be placed on them to make distorted perceptual judgments, the experimenter is using a model of natural conformity in the expectation that its results can be generalized. An experimenter who memorizes lists of nonsense syllables and measures the decline in recollection of them is using a model of verbal learning and memory.

Key Ideas Box 2.4: Selecting a setting for research

Experimenters should give thought to the best setting for carrying out their research. Experimenting in the natural *target setting* of interest, though sometimes difficult, maximizes generalizability, minimizes the tendency of subjects to modify responses because they are "in an experiment," and allows all natural behaviors to occur.

The second option is to use a *model setting*. Model settings differ in the extent to which they resemble the natural setting. Sometimes settings that do not obviously resemble the target setting will, nevertheless, reveal the same functional relationships. This must be checked by the usual methods of evaluating generality.

The range of potential models is great, and an appreciation of it will be gained by reading the sample research later in the book. Obviously, model settings can range from close imitations of the target setting to settings that are thought to be similar to it only in a functional sense.

What do we mean by "similar in a functional sense"? Two behaviors may appear to be quite different, but the functional relationships underlying them may be alike or even identical. When Reedy the rat presses a bar for food, he may be executing a behavior functionally identical to little Georgie's saying "Mik" at the table when he wants more milk. If Janie's parents were alcoholic, she may engage in self-destructive behavior that merits her being considered a "dry alcoholic." Without drinking at all, Janie might imitate the essential self-destructive behavioral characteristics of the alcoholic parents.

SELECTING THE SUBJECTS

Subjects from the target population

Here again, you can use subjects directly from the target population. If your question is about people, you can use human subjects; if it is about a given nonhuman species, you can use subjects of the kind involved directly in the question. This has the obvious advantage of strengthening the generalizability of the findings.

Model subjects

If you cannot sample from the target population, you may have to use model subjects, in particular animal subjects. The selection of animal subjects will depend on logistical considerations and on certain strategic decisions.

SELECTING CONVENIENT SUBJECTS SUCH AS RATS. If you are doing an experiment in which you seek broadly general, basic behavioral principles, the nature of the animal subjects may matter little. If you look for the most general laws of behavior, any organism displaying behavior will serve your purpose.

It is a misconception to suppose that animal subjects must be *representative* in the sense of being like human subjects or like higher mammals to a special degree. Lockard (1968) has argued against the use of rats as subjects in psychology on the grounds that rats are not sufficiently representative. He suggests that the use of rats is a "bad habit." But general laws are often discovered on species that are convenient, and by no means ordinary or typical.

When geneticists spent many years studying the genetics of the tobacco mosaic virus, they were dealing with a rather exceptional kind

Key Ideas Box 2.5: Selecting subjects

Subjects should be selected according to the best possible *sampling procedures* in order to avoid biased sampling. Furthermore, subjects should be selected to maximize generality instead of taking it for granted that certain familiar types of subjects (such as white rats or college sophomores) will be used.

First, ask whether subjects from the target population can be used. If not, what apt model subjects should be used? If you seek the most general principles of behavior, perhaps any behaving organism will do. More often, one type of subject will model the target population better than another. For example, rats seem to be poor models for the study of alcoholism, since they are reluctant to drink alcohol.

of organism. But it happened to be convenient for the procedures they wanted to use. Many Nobel Prizes were won, and ultimately the genetic code was understood as a result of such work. Neurophysiologists made the first direct measures of conduction in individual nerve fibers on a downright freakish preparation, the giant axon of the squid. With the geneticists and the neurophysiologists the preparation used was technically convenient, and the principles sought were of the most general kind.

SELECTING APT MODEL SUBJECTS. Another approach is to attempt to select model subjects that are particularly representative with respect to some property of the target population. Physiologists commonly do this sort of thing. If they are interested in kidney function they will use humans if this is not threatening. If they cannot use human subjects, instead of just taking the subjects at hand, they will find an animal model that has a kidney as similar as possible to that of humans. When psychologists insist on using primate subjects for potentially traumatic experiments, they are taking an attitude similar to that of the physiologists. When students of alcoholism move away from the use of rats and monkeys, which species display a singular lack of enthusiasm for alcohol, and shift to a special kind of pig which takes readily to drink, they are selecting apt subjects.

Investigators like those mentioned are usually studying specific complex phenomena that may only exist in a limited number of species. Their strategy is different from that of experimenters interested in broadly general basic principles.

Study Questions

1. What is generality of data?
2. Distinguish between population and sample.
3. What are the bases for judging that data are general?
4. Explain the following kinds of sampling:
   a. representative
   b. random with replacement
   c. random without replacement
   d. stratified
   e. catch as catch can
   f. biased
5. To what extent do most samples in psychology meet the criteria of true randomness?
6. What is the principle of parsimony and how does it relate to generality?
7. Why not design experiments to affirm null statements?
8. Explain experimenter effects and experimenter bias.
9. What does it mean to say an experiment was double blind?
10. What are the advantages and disadvantages of doing experiments in the target setting?
11. What does it mean to experiment in a model setting?
12. Taking the discussion of selecting subjects into account, discuss the issue of whether rats should be used as subjects in psychological experiments.

Exercise

1. Go to the library and find an issue of *The Journal of Experimental Psychology.* Make a list of the species studied by authors in that issue. To what extent are the authors assuring generality? Do the same thing for an issue of *The Journal of Comparative and Physiological Psychology* and for an issue of *The Journal of Personality and Social Psychology.* While you are looking at those journals, notice whether authors tried to assess generality across dimensions other than species.

3 DESCRIBING DATA

Important functional relationships may be concealed in a mass of raw data. A critical step in manipulating data is reducing them to a form in which these relationships are exposed and susceptible to evaluation for such properties as reliability and generality. The data must be reduced and presented in a way that brings home to the experimenter as well as to others the full import of the findings. Here we will discuss various useful procedures for so reducing and representing data.

Graphing data

Let's assume that you have gathered a set of data. It is most unlikely that the numbers obtained from different subjects or from the same subjects at different times will all be identical. Instead, there will be a degree of scatter of the scores. We call the scattered set of scores a *distribution.* Experimenters describe distributions according to their geometry, so it is important to start by describing the relationship between geometry, numbers, and equations.

Each number obtained from experimental observations can be represented as a location or point on a line, as in the following illustration:

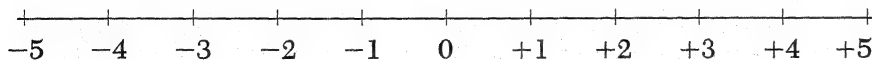
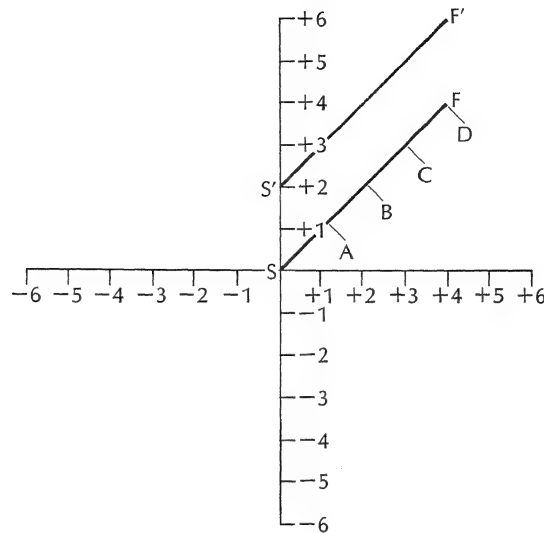



FIGURE 3.1 Graph of the equations $V = H$ (line S–F) and $V = H + 2$ (line S′–F′). Adding a constant changes the point of interception.

Other, more complex figures can be represented on a system of Cartesian coordinates. In experimental work, we place values of the *independent variable on the horizontal axis* and values of the *dependent variable on the vertical axis.*

THE SLOPE OF A LINE

Later, we will use the concept of the slope of a line. To visualize the meaning of this concept, compare the line formed when $V = H$ to the line formed when $V = 2H$, as in Figure 3.2. The line climbs much more abrubtly when $V = 2H$. If you were a beginning skier, you would much rather ski down the $V = H$ line, because the slope is steeper when $V = 2H$.

The notion of the slope of a line is easy to see geometrically. We need to have an algebraic representation of it. Form a little right triangle on the bottom side of the line. Now form a fraction with the size of the vertical side of the triangle on top and the size of the horizontal side of the triangle on the bottom. This will give the slope of the line. For line SF, the fraction is 1/1, and for line S′F′, the fraction is 2/1. You can see that the larger this fraction, the steeper will be the slope of the line. Notice that what you are doing with this maneuver is equivalent to dividing V by H. For example, take the last point on line S′F′. Its values are $V = 8$ and $H = 4$. The fraction $8/4 = 2$, and so does the slope, as we determined it above, equal 2.

You cannot always get the slope of a line by dividing V by H. If we take line S′F′ in Figure 3.1, the slope calculated by the method of little

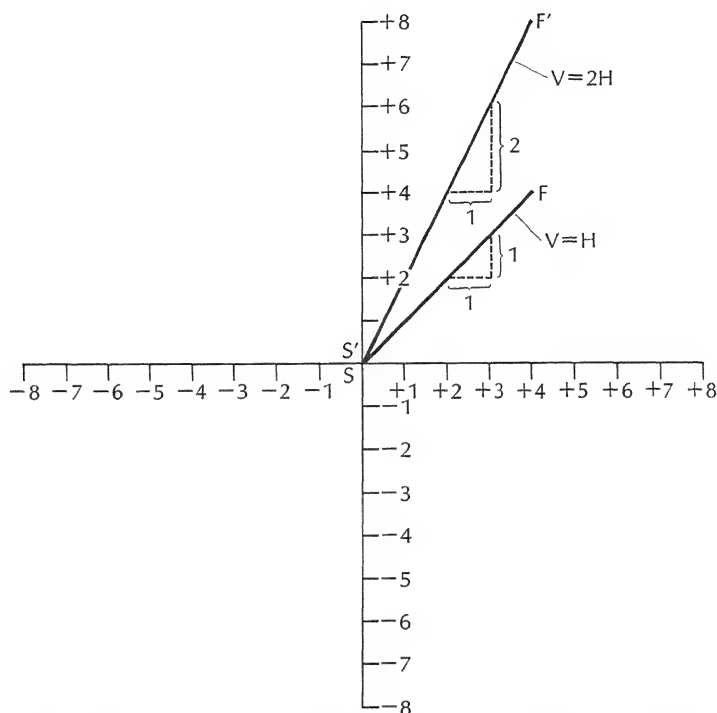



FIGURE 3.2 The slope of a line is its steepness. It is reflected in the ratios of the sides of the right triangles projected below the line. The ratio for line S–F is 1/1; for line S′–F′ it is 2/1. See text for details.

right triangles is equal to 1. But if you take, for example, the last point on the same line, $V = 6$ and $H = 4$, the result is not 1 ($6/4 \neq 1$). What you have to do is correct for the V intercept by subtracting the intercept value from V before forming the fraction to represent the slope. So, subtracting 2 from 6 gives 4, and $4/4 = 1$, the correct value. A general statement of the slope is

$$\text{Slope} = \frac{(\text{Vertical value}) - (\text{Vertical intercept})}{(\text{Horizontal value})}$$

GEOMETRIC REPRESENTATION OF DISTRIBUTIONS

The set of numbers gathered in an experiment can be represented geometrically on a system of Cartesian coordinates. A simple representation of a distribution can be given by dividing the scores into classes, which are ranged along the horizontal axis, and then representing the frequencies of scores of the classes on the vertical axis. The result is called a histogram. Several histograms are shown in Figure 3.3.

The distribution represented by the histogram at the top (Figure

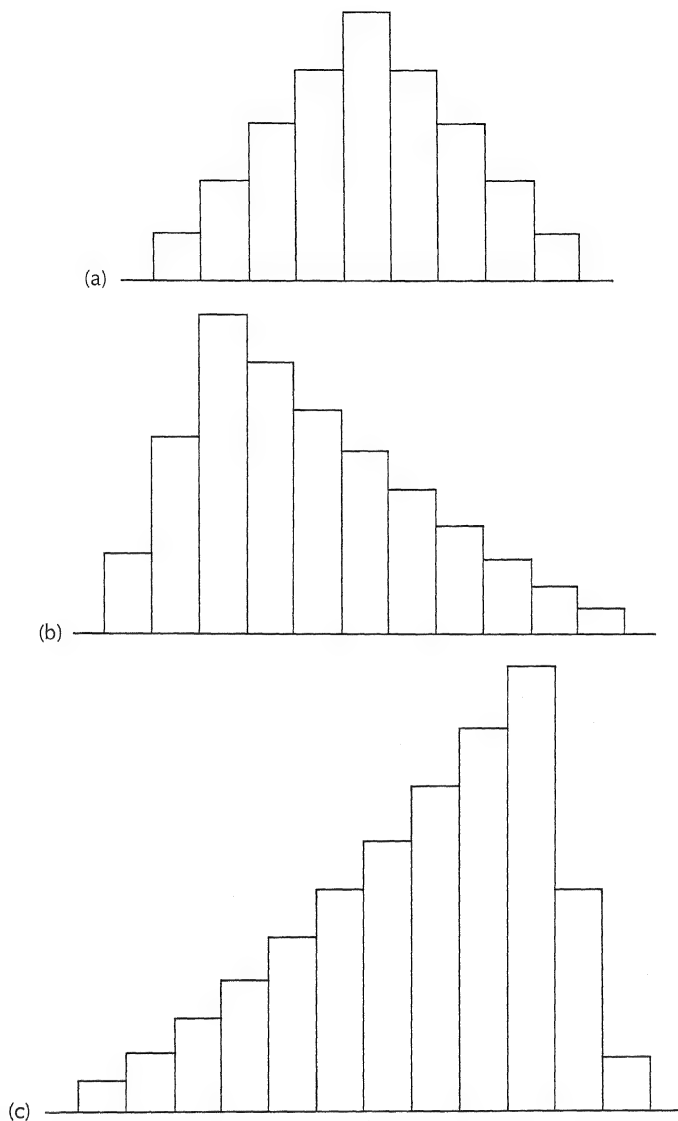



FIGURE 3.3 Histograms showing (a) symmetrical distribution; (b) Positively skewed distribution: note that it has a long, thin tail to the *right;* and (c) negatively skewed distribution: note that it has a long, thin tail to the *left.*

3.3a) is symmetrical, or evenly distributed about its center. The other two distributions are *skewed,* or unbalanced to one side. A distribution is said to be skewed in the direction of its long, skinny tail. Thus, distribution b is positively skewed (tail at the right, toward higher positive values on the H axis) and distribution c is negatively skewed.

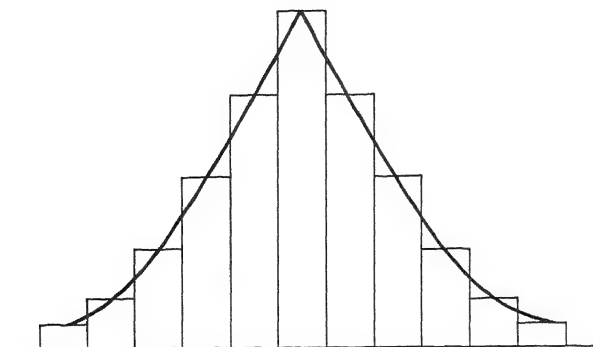



FIGURE 3.4 Formation of a curve by connecting the centers of each bar on a histogram.

If the number of categories of a histogram is increased, the bars become more and more slender. In principle the process of increasing categories and producing increasingly narrow bars could go on until the distribution would appear to be an area bounded by a smooth curve. A histogram thus approaches a smooth curve as the number of categories increases. Distributions are often expressed as curves rather than histograms.

Another way to create a curve instead of a histogram would be simply to place a dot on the system of coordinates representing the central value of each bar on the histogram and then connect the dots together. Figure 3.4 shows how this might be done.

Like the histogram, this representation of the distribution can vary in a number of ways that make visible certain properties of the underlying data. Whether for histogram or curve, the main features of interest are: (1) The location of its center, which is called its *central tendency.* (2) The degree of its scatter around the center (that is, is it narrow or is it broad). This is called its *dispersion,* and indicates how variable the measurements are. (3) Its *skewness* or tendency to be shifted to the right or the left. (4) Its *kurtosis,* or the extent to which it is peaked or flat. All of these geometrical properties represent more or less important characteristics of the underlying data.

Sometimes we need to have a geometrical representation of the data and at other times we need a numerical representation. Both of these help us to understand our data and to present them to others.

The normal or Gaussian distribution

The "normal" distribution will crop up in various contexts later, so let us describe it here. Figure 3.5 shows a "normal" distribution. It is bell-shaped and symmetrical about its high point. The "normal" curve is also called a Gaussian curve, after the great mathematician Karl Friedrich Gauss.

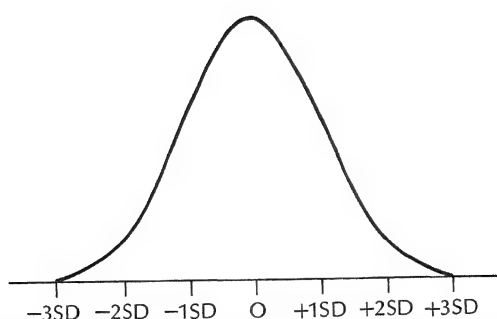



FIGURE 3.5 A Gaussian distribution. The markings on the horizontal axis indicate standard deviations (SD), which are explained later.

Approximations to the Gaussian distribution in sets of observations occur often, especially when many random factors are each having a small influence on the observations, some positive and others negative in direction. This means that the data from well-controlled experi-

Key Ideas Box 3.1: Describing data in numbers, equations, and graphs

Data usually take the form of numbers. For many purposes, numbers, equations, and graphs are interchangeable. For a particular set of data, the meaning may be clearer in one of these forms than in another. Besides numbers, graphs are commonly used to express data. Values of the independent variable go on the horizontal axis and values of the dependent variable go on the vertical axis. Information may be gleaned merely from looking at a graph. Sometimes numerical or algebraic indicators of general properties of the graph are used. An important one is the slope of a line. Information on how to determine the slope of a line is given in the text.

When data are gathered, they generally vary (are "distributed") around some central value. The resulting distribution may be plotted in a bar graph (histogram) or a curve. There are many forms of distribution. Some are positively skewed (have a long tail to the right); others are negatively skewed (have a long tail to the left); and others are symmetrical. The curve of *normal* or *Gaussian distribution* is symmetrical and bell-shaped (see Figure 3.5). Approximations to the Gaussian distribution occur often in nature, especially when deviations from the central value are due to random influences of many small factors.

ments, from which constant errors have been taken out, might well be expected to take on a Gaussian distribution. For this reason, scientists and mathematicians dealing with errors of measurement have commonly assumed observations to follow a Gaussian distribution.

Numerical description of distributions

An experimenter can easily miss features of a distribution if he tries to look at it as a whole. We have a number of measures to summarize major characteristics of a distribution. These include measures that indicate where the center of the distribution is located, how scattered the scores are around that center, the degree to which the distribution is flat or peaked, and the extent to which it is symmetrical about its center rather than shifted to one side. We will discuss the most important of these measures here.

MEASURES OF CENTRAL TENDENCY

Experimenters probably attend more to the central tendency of their data than to any other of its features. This stems from the belief that the measure of central tendency reflects the true measure better than anything else. We will treat four different measures of central tendency here. These are the arithmetic mean, the geometric mean, the median, and the mode.

The arithmetic mean

The arithmetic mean is the sum of all the observed values divided by the number of observations. This is the familiar "average" taught in grammar school. The mean of 2, 3, and 4 is simply 2 + 3 + 4 divided by 3 (= 3). Stated in general terms, the arithmetic mean is

$$\frac{\sum_{i=1}^{n} X_i}{n}$$

What do these symbols mean? Since they will be useful later, let's go through them step by step. The Σ means "the sum of," and the symbol X means the set of observed values, from first to last. The letter X is used to indicate an observed value, and if there are numerical subscripts, such as X_1, X_2, and X_3, they indicate *which* observation we have in mind, the first, second, third, and so on. When a letter such as the i in the formula is used as a subscript, it indicates that we have in mind the scores in general, no one in particular. So ΣX_i says simply "sum the scores." On top of the Σ we place a letter indicating when to stop summing. When the letter n occurs at the top of Σ, it means to sum until the last score is included. At the bottom of the Σ we indicate

where to start summing. The indication in the formula says to start at the first score. So the top part of the formula says "sum the scores, starting from the first one and going all the way through the last one." The bottom of the formula says to divide by the number of observations.

The arithmetic mean is surely the most widely used measure of central tendency in science. It is so widely used that when someone refers to the "mean" without indicating what kind of mean, we take it for granted the arithmetic mean is intended. Accordingly, we will drop the adjective "arithmetic" in most cases, and understand the term "mean" to refer to the arithmetic mean unless otherwise stipulated.

Individual scores typically vary around any measure of central tendency, including the mean. If the mean of 2, 3, and 4 is 3, two of the scores deviate from the mean; only one of them is right on it. If you determine how much the scores deviate from the mean (in this case 2 deviates by -1 and 4 deviates by $+1$), the sum of the resulting "deviation scores" will always equal zero. The deviations above it always equal the deviations below it.

Mathematicians often get rid of $+$ and $-$ signs by squaring numbers. A positive number times a positive number yields a positive number, and a negative number times a negative number also yields a positive number. Hence any number multiplied by itself will result in a positive number. In describing and analyzing data, we commonly work with squared deviation scores. The mean has a special property with respect to these squared deviation scores. It is this: the sum of the squared deviations from the mean is equal to or less than the corresponding expression with any other measure of central tendency. Put another way, the sum of the squared deviations from the mean is at a minimum.

The geometric mean

The geometric mean is not used often with psychological data, but you will encounter it from time to time. For example, it is typically used on data obtained by the method of magnitude estimation in psychophysics. It is particularly useful when a distribution is skewed. To get the geometric mean, you find the product of all the n scores, then take their nth root. Thus, in comparison to the method of getting the arithmetic mean, we substitute multiplication for addition, and we substitute taking a root for the operation of division. The geometric mean of 2, 3, and 4 is

$$\sqrt[3]{2 \times 3 \times 4} = \sqrt[3]{24} = 2.9$$

It would be very laborious to calculate the geometric mean in this way for large numbers of observations. But we know how to do it indirectly and quite simply by using logarithms. The operation of

multiplying numbers is equivalent to the operation of adding the logarithms of the numbers, and the operation of taking the nth root of a number is equivalent to dividing its logarithm by n. It is easy to convert numbers to logarithms, to add, and to divide, so the geometric mean is not hard to find. We use the geometric mean in the same situations in which we use logarithmic transformations. We deal with logarithmic transformations later.

The median

The median is also useful for skewed distributions. The median of a number of observations is that value which has an equal number of observations greater than and less than itself. It is the point above and below which half of the observations lie. If there are an uneven number of observations, we find the median by placing the scores in order of magnitude, enumerating them, and taking the middle item. Since there is no middle score when we have an even number of observations, we use the arithmetic mean of the middle two observations as the median in such cases.

When the distribution of scores is symmetrical, the mean and median are equal. They deviate from each other when the distribution is skewed because the mean is highly sensitive to extreme scores, whereas the median is insensitive to them. For example, the median of 2, 3, and 4 is 3, and so is the mean. But the median of 2, 3, and 1000 is also 3, whereas the mean is 335.

Key Ideas Box 3.2: Describing central tendency

Data tend to vary around a central value. Several different numerical representations of that value are in common use. The most widely used, the *arithmetic mean,* is equal to the sum of scores divided by the number of scores. The *geometric mean* is the antilog of the arithmetic mean of the logarithms of the scores. It is used for highly skewed distributions because it gives less weight to extreme values than does the arithmetic mean. The *median* is the middle value of the scores in a distribution (if an even number of scores, it is the arithmetic mean of the middle two values). It, too, is relatively unresponsive to extreme values, but it only takes into account a few of the data. The *mode* is the most frequently occurring score. It represents the distribution accurately. However, it can only be used when there are many scores repeated. This is not likely to occur with small samples, so the mode is not often used in experimental work.

The mode

The mode is the most frequent score value in a distribution, or the most frequent class of score values. It is rarely used as a measure of central tendency in experimental psychology, because sample sizes are commonly not great enough to provide large numbers of repeated scores.

THE DISPERSION OF DISTRIBUTIONS

A second feature of distributions we need to describe is their scatter around the center, their *dispersion.* We do this by using the deviation scores about the mean as an index of dispersion. The sum of the deviation scores about the mean always equals zero, so it provides a poor basis for indicating dispersion. We need to get rid of the + and − signs. This can be done by simply dropping them. The resulting numbers are the absolute deviations from the mean. Customarily we take the mean of the absolute deviations, so that we have an indication of the average amount of deviation per score.

The *mean absolute deviation* is an adequate measure of dispersion, but happens to be seldom used. The second method of getting rid of the directional signs on numbers, that of squaring, has been adopted as the preferred method. We take the deviation from the mean for each score in the distribution and square it. Next we take the mean of these squared deviations. This gives the *mean squared deviation.* Symbolically we can express it as follows (d = deviation):

$$\text{Mean squared deviation} = \frac{\sum_{i=1}^{n} d_i^2}{n}$$

This says: "Take the sum of the squared deviation scores and divide by the number of scores." You can see that the formula is the one for the arithmetic mean, but deviations scores are used instead of raw scores. To represent how the deviation score is derived, we could rewrite the equation this way:

$$\text{Mean squared deviation} = \sum_{i=1}^{n} \frac{(X_i - M)^2}{n}$$

with X representing "the scores in general," M representing the mean of the scores, and n representing the number of scores. Another name for the mean squared deviation is the *variance.*

Since the variance has been inflated by the operation of squaring, we may want to deflate it again by taking its square root. The square root of the variance is called the *standard deviation.*

$$\text{Standard deviation} = \sqrt{\frac{\sum_{i=1}^{n} d_i^2}{n}}$$

In practice, both the variance and the standard deviation are used as measures of the dispersion of distributions.

If a distribution has the symmetrical, normal configuration, about 68 percent of the distribution will be included between plus and minus 1 standard deviation, about 95 percent will be included between ± 2 standard deviations, and about 99.73 percent will be included between ± 3 standard deviations.

There are many measures of the dispersion of a distribution. The ones already described are most commonly used. Experimenters often use a cruder measure, the *range*, as a rough indicator of dispersion. The range is merely the difference between the lowest and the highest score in the distribution. It is a cruder measure than the three already described because it is only influenced by two of the scores in the distribution, whereas all of the scores are taken into account by the mean absolute deviation, the variance, and the standard deviation.

Corrections and transformations of data

CORRECTIONS

Raw data are not necessarily the best representation of reality. Errors of measurement are included in any recorded datum. These errors are of two basic kinds: variable errors and constant errors. We will deal with variable errors later. The most usable datum may be one that has been

Key Ideas Box 3.3: Describing dispersion

It is important to have a number to describe the extent to which scores spread out around the central tendency of a distribution. Two distributions with markedly different degrees of spread can have the same center. The *variance* reflects dispersion. To get the variance, we subtract each score from the mean, square the resulting deviation scores, then take the mean of the squared scores. The *standard deviation* is the square root of the variance. A rough measure of dispersion is given by the *range*, which is the difference between the lowest and the highest score. Since it only takes two of the scores into account, it is less informative than the variance and standard deviation.

corrected for various constant errors. This is true for hard sciences as much as for psychology. For example, observed measurements of the positions of the sun and stars are thoroughly contaminated by constant errors due to such factors as the capacity of air to bend light. The sun can still be seen when it is below the horizon. Consequently, astronomers must introduce corrections before treating measurements as basic data.

Psychologists often correct data according to "baseline" scores obtained in previous behavioral observations. For example, Besley and Sheridan (1973) trained rats to do diametrically opposed visual discriminations, one with each cerebral hemisphere. They were interested in the effects of tiny areas of brain damage made in one hemisphere of the brain. The idea was to see whether the balance of evenly matched opposing discriminations mediated by opposite hemispheres might be shifted by small areas of damage, not behaviorally detectable by ordinary methods. They could have simply used the numbers of errors made by the damaged hemisphere after surgery. But even before surgery there were sizeable variations in the numbers of errors. The occurrence of a large number of errors after surgery on the part of an animal that had made a large number of errors prior to surgery would not mean as much as a similarly large number of postoperative errors by an animal that had done very well before surgery. Consequently, Besley and Sheridan (1973) subtracted for each animal the number of errors prior to the operation from the number of errors after the operation. The resulting "corrected" or "difference" scores gave a more meaningful measure.

LINEAR TRANSFORMATIONS

Whether or not we correct data for known sources of constant errors, we may want to transform them. Linear transformations are done when the experimenter wishes to change the zero point of a set of measures, or to change the units of measurement. One may want to do this in order to make computations easier, as when the data include both positive and negative values. For example, adding a constant +7 to the following set of numbers converts them to uniformly positive numbers:

$-4, 13, -3, 4, 7, -7$ (original)
$3, 20, 4, 11, 14, 0$ (transformed)

Linear transformations are done by adding, subtracting, multiplying, or dividing all the scores by the same thing. The common procedure of going from one unit of measurement to another through multiplying by a constant number is an example. If I want to convert a certain set of

measurements from feet to inches, I multiply all the measurements in feet by 12.

Linear transformations permit us to convert a distribution of scores in such a way as to achieve a desired mean and standard deviation. Multiplying a set of numbers by a constant will increase the standard deviation by the same ratio. This means that doubling each score doubles the standard deviation, tripling each score triples the standard deviation, and so on. This operation will also increase the mean by the same ratio. If you want to increase the mean without affecting the standard deviation, you do this by adding a constant to each score. The mean will be increased by that same constant, but the standard deviation will be unchanged. An application of these maneuvers to the Z_i score is discussed later in this chapter.

NONLINEAR TRANSFORMATIONS

There are various types of nonlinear transformation that experimenters make on data for various special purposes. The *logarithmic transformation* is one of the most common. The logarithm of a number is the exponent indicating the power to which another number must be raised in order to get a result equal to the original number. The "other number" to which the exponent is applied is called the *base* of the logarithm. Familiar bases are 2, *e* (2.71828), and 10. Logarithms to the base *e* are called *natural* or *Napierian logarithms.* Logarithms to the base 10 are called *common logarithms.* We usually indicate which base we are using by a subscript; thus $\log_2$ means that the base is 2, and similarly $\log_e$ and $\log_{10}$ indicate logarithms to the base *e* and to the base 10, respectively. If there is no subscript, this means that the base is 10.

An impression of the relationship between logarithms and their corresponding numbers (called *antilogs*) can be obtained from the following list:

Number	Equivalent in terms of 10 raised to a power	Common logarithm
10	10^1	1.0000
100	10^2	2.0000
1000	10^3	3.0000
10,000	10^4	4.0000

The logarithm is merely the power to which 10 is raised. Notice that the logarithm grows much more slowly than the number itself. In fact, the numbers in the list are in *geometric progression*—each successive number is a constant multiple of the preceding number. In this case, each successive number is 10 times the one preceding it. The list of

logarithms is in *arithmetic progression.* This means that each successive logarithm adds a constant unit (in this case, one) to the one preceding it.

Because of this tendency to shrink a geometrically expanding series of numbers, it is often useful to convert raw data into logarithms of the original scores. It sometimes happens that the size of the increments of change of the dependent variable is not uniform, but varies systematically with increases in value of the independent variable. This occurs, for example, when each successive value of the dependent variable is a constant *percentage* of the preceding value. As a practical matter, such functions grow so rapidly that it can be difficult to plot them on a graph. Furthermore, the resulting distribution of scores is skewed. When a distribution is skewed, the arithmetic mean does not represent it well, and assumptions of certain statistical tests cannot be met. Hence it is useful to convert from the scores to their logarithms.

A study done in the laboratory section of my experimental psychology class illustrates an application of logarithmic transformation. People were asked to assign numbers rating the personal value of their various bodily parts—eyes, ears, nose, and so forth. The numbers could be as large or as small as they wanted to use, but should consistently reflect the relative values of the parts. (Later, in Chapter 8, you will learn that this is the psychometric method of magnitude estimation). One participant used numbers like 100,000 for certain parts. Most subjects stayed with numbers like 100 and 60. If the arithmetic mean of these scores were used, a single subject using very large numbers could radically distort the measure of central tendency. One way to reduce the impact of the varying sizes of numbers chosen is to use a logarithmic transformation. The common logarithm of 100 is 2 and of 100,000 is 5. Thus the extreme weightiness of large numbers is made manageable by logarithmic transformation. This is, of course, essentially the same as using the geometric mean. Many other examples of the use of logarithmic transformations will be found in Chapter 8.

A second transformation of considerable usefulness is the z or *standard score* transformation. This transformation makes it possible to compare scores that were originally not in comparable units. For example, if a man is 6 feet tall and weighs 200 pounds, he is taller and heavier than the average man. But is he heavier to a greater extent than he is taller? This is impossible to answer with the scores in their original form, but can be answered when they have been transformed into z scores. The formula for z is

$$z = \frac{\text{(Score)} - \text{(Mean of scores)}}{\text{(Standard deviation)}} = \frac{(X_i - M)}{\text{SD}}$$

The deviation from the mean is expressed in units equal to the standard deviation. Thus, each score is expressed in terms of the relative

proportion of scores in that distribution that deviate so much from the mean. It is like saying that the man is taller than 90 percent of the people but heavier than only 75 percent, thus he is taller to a greater extent than he is heavier. (These percentages are, incidentally, entirely fanciful).

The z transformation as described will provide a distribution with a mean of zero and a standard deviation of one. Since this results in both positive and negative numbers, computational ease[2] leads many investigators to go a step further and obtain a distribution with a preassigned mean and standard deviation. Since adding a constant to each score increases the mean by that same constant and multiplying each score by a constant increases both the mean and the standard deviation by that constant factor, it is easy to achieve the desired mean and standard deviation. For example,

$$Z_i = 50 + 10\,\frac{(X_i - M)}{\text{SD}}$$

will give a mean of 50 and a standard deviation of 10. This is because the last term in the equation is equal to the original z transformation, in which the mean was zero and standard deviation, 1. Adding 50 to each score increases the mean by 50. The mean was zero, so multiplying it by 10 does not change it. The standard deviation was 1, so multiplying it by 10 results in a product of 10.

Describing degrees of relationship

Experimenters often want a measure of *how much* relationship exists between two variables. We have a number of ways to describe the degree of a relationship. Consider what happens when two variables, H and V, are so perfectly related that $V = H$ for all values. Going back to the beginning of this discussion of describing data, remember that the geometric counterpart of this equation is a straight line. It would be a straight line with a different slope if $V = 2H$, and it would be a straight line with a different slope and a different intercept if $V = 2H + 2$.

Sometimes two variables are related very intimately, but when values of one of them rise, values of the other decline. Though the relationship is *inverse,* it is just as useful in predicting values of one variable from values of the other. I once knew a psychologist who was working for a manufacturer, trying to devise a test to predict certain mechanical aptitudes. The authorities at the firm felt that skill in algebra was important to the type of mechanical job they had in mind.

[2]There are psychological considerations as well. How would you feel if your score on a test came out to be a negative number?

Key Ideas Box 3.4: Correcting and transforming data

It is not always best to work with raw data. Data may be improved by *correcting* them for constant errors. Constant errors are errors that influence data consistently in a given direction. Data may be corrected by subtracting known values, for example scores of individuals on a pretest, from the scores. There are also a variety of transformations of data that aid us in handling and understanding data. *Linear transformations* (adding, subtracting, multiplying, or dividing all scores by the same thing) change the zero point or the units of a set of measures. They can make computations easier, for example, by eliminating negative values.

Logarithmic transformations (converting each number to its logarithm) can make skewed distributions more symmetrical by shrinking extreme scores. This may be useful in meeting the assumptions of certain statistical tests and in making it easier to plot all the scores on one graph.

Standard (z) scores make it possible to compare data that were originally not in comparable units (for example, to answer the question: am I more exceptionally tall than I am exceptionally bright?). The z score indicates what proportion of the population falls above or below you on a given dimension. The Z_i transformation further permits giving the distribution any desired mean or standard deviation. The formula is

$$Z_i = \text{Desired mean} + \text{Desired SD}\,\frac{(X_i - M)}{\text{SD}}$$

He included measures of algebraic skill in a battery of tests, and found that the more skilled people were at algebra, the worse they were at the mechanical task. He went to the manufacturers and said, "Don't hire anybody who is good at algebra." The psychologist was right, but the manufacturers got rid of him.

If a relationship is inverse, as when $V = \frac{1}{H}$, the geometrical representation is still a straight line, but its slope is downward instead of upward. The slope is negative instead of positive. The fact that some relationships can be represented by the plot of a straight line has been used to produce a measure of degree of relationship. Of course, real data will seldom come out to fit anything so simple as a straight line. There will be variability around H and there will be variability around V. Even assuming that the true measure of things would be a straight

line, error of measurement would likely lead to scatter around the straight line. Consequently, we are not deterred when we plot individual H and V scores against each other and get a kind of football-shaped scatter of points. Instead, we call it a *scatter diagram* and find the straight line that fits it best. There are many definitions of "best fit," but the one generally used is the straight line for which the sum of the squared deviations from the line is at a minimum. The best-fit line is called a *regression line* and its equation is a *regression equation.* The slope of the regression line indicates the degree of relationship between the two variables.

This provides you with an intuitive notion of how the degree of a relationship can be represented geometrically and algebraically. In practice, correlation coefficients are used. The most widely used one is the *Pearson product-moment correlation coefficient,* symbolized as r. The maximum positive relationship is indicated by a value of +1.00, absence of relationship yields r = 0.00, and a maximum negative (inverse) relationship provides a value of −1.00. Intermediate degrees of relationship produce intermediate values.

Pearson's r is used for data measured on at least an interval scale,[3] and it assumes that the data can legitimately be represented as a straight line and that the standard deviations of values of the two variables are equal. A definitional formula of r is

$$r = \frac{\sum_{i=1}^{n} h_i v_i}{N(\mathrm{SD}_h)(\mathrm{SD}_v)}$$

where h and v are deviation measures from the H and V means (that is, $\mathrm{H}_i - M_{\mathrm{H}} = h_i$ and $V_i - M_v = v_i$); N = the number of individuals measured; and SD_h and SD_v are the standard deviations of the two distributions.

A computational formula is

$$r = \frac{N\Sigma HV - \Sigma H \Sigma V}{\sqrt{N\Sigma H^2 - (\Sigma H)^2}\sqrt{N\Sigma V^2 - (\Sigma V)^2}}$$

Translated into words, this says

a. $N\Sigma HV$. Multiply the V value by the H value for each of the individuals, add up the resulting products, and multiply the sum by the number of individuals.

b. $-\Sigma H \Sigma V$. Add up all the H values, then add up all the V values, and multiply the resulting sums. Subtract the result from the result of step a.

[3]Types of scales are explained in Chapter 5. The interval scale requires that scaled items be (1) reliably identified, (2) placed in order on the scale, and (3) spaced according to unit intervals that are equal regardless of what part of the scale is used.

c. $\sqrt{N\Sigma H^2-(\Sigma H)^2}$. First, for $N\Sigma H^2$ square each H value and sum the results, then multiply by the number of individuals. For $-(\Sigma H)^2$ add up the H values, square the result, and subtract it from $N\Sigma H^2$. Take the square root of this difference.
d. $\sqrt{N\Sigma V^2-(\Sigma V)^2}$. Do the same things to the V scores that you did to the H scores in step c.
e. $\sqrt{N\Sigma H^2(\Sigma H)^2}\ \sqrt{N\Sigma V^2-(\Sigma V)^2}$. Multiply the result of step c by the result of step d.
f. Divide the result of step b by the result of step e.

There are many other correlation measures (for easily used descriptions see Siegel, 1956, and Bruning & Kintz, 1968), but a second one that is very often used is Spearman's ρ (rho). Experimenters use this to find the degree of relationship between two sets of data that are on an ordinal scale. A formula is

$$\rho = \frac{1-6\,\Sigma D^2}{N(N^2-1)}$$

Put in words, the formula is as follows:

a. $6\,\Sigma D^2$. Subtract the H from the V value for each individual (observation), square each difference (D), add the results, and multiply by 6.
b. $N(N^2-1)$. Square the number of individuals, subtract 1, and multiply the result by the number of individuals.
c. Divide the result of step a by the result of step b.
d. Subtract the result of step c from 1.

CORRELATION VERSUS CAUSATION

No matter how strong the correlation between two things might be, it does not show that one causes the other. For example, there have been reports of a relationship between the amount of coffee a person drinks and the likelihood of a heart attack. The effect is not due to caffeine, because the consumption of tea, which has at least as much caffeine as coffee, has no such relationship to heart attacks. There *may* be some other factor in coffee that encourages heart attacks, but then again there may not. People who push themselves very hard, working late into the night, often bolster themselves with many cups of coffee. The relationship between coffee drinking and heart attacks might be simply due to the fact that hard-driven people are simultaneously more likely to drink a lot of coffee and to have heart attacks. In this case the relationship between coffee and heart attacks would be correlational but not causational.

The search for controlling variables is quite distinct from the mere search for correlations. We get knowledge of controlling variables from

Key Ideas Box 3.5: Correlation, causation, and describing degrees of relationship

When two variables are related, the relationship may or may not be causal. We cannot conclude from the presence of a relationship or its degree that the independent variable caused the variations in the dependent variable. For example, it has been observed that the stock market varies along with skirt lengths, but the one does not cause the other.

Whether causal or correlational, relationships must be described. There are many measures to describe degree of relationship. Pearson's product-moment correlation coefficient *(r)* is widely used for data on an interval scale when the standard deviations of the two variables are about equal. The formula for r is given in the text.

For rank-order data, Spearman's ρ may be used. See text for details.

experimental work. For example, the first evidence suggesting that smoking was hazardous to health was correlational. The likelihood of certain diseases, such as cancer of the lung, was substantially related to the number of cigarettes smoked. This evidence, like the present evidence on coffee and heart attacks, was not enough to permit conclusions about an actual influence of cigarette smoking on cancer. Later, however, it was possible to show that lung cancer could be induced experimentally in dogs by introducing cigarette smoke into their lungs.

Should we use grouped data in psychology?

Psychologists have argued a great deal about whether they should rely on functions derived from group averages. Science has always used averages. The presumption is that any given observation is composed of two parts, a true measure and error of measurement. By taking averages, errors upward cancel errors downward, and a good estimate of the true measure can be obtained.

However, the argument is not over the use of averages. It is over the issue of averages *across individuals.* Fundamentally it boils down to a question of whether behavioral variations from one individual to another can be regarded as errors of measurement. Most pyschologists are reluctant to regard them in this way. But averaging of grouped data implies that they are errors of measurement.

Figure 3.6 shows how different individual functions can be from

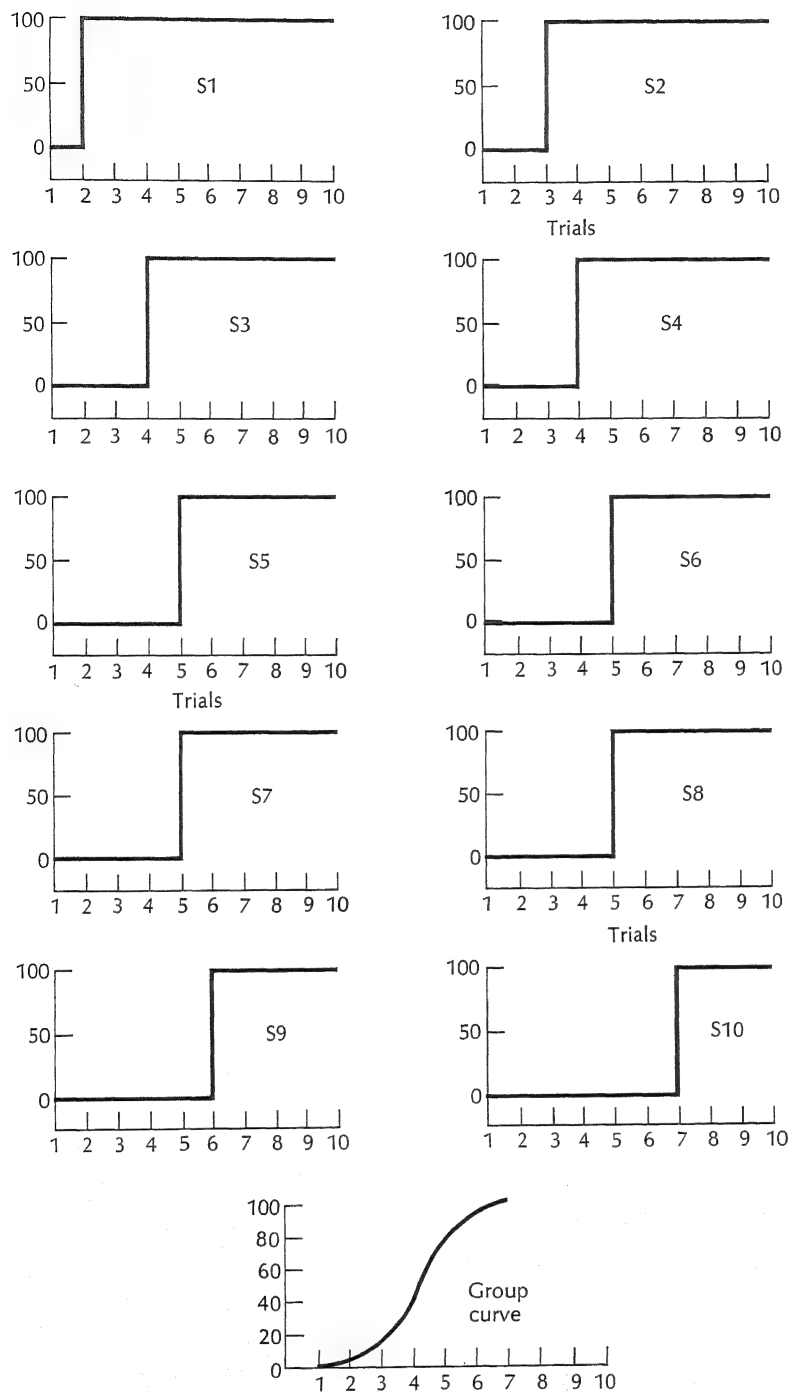



FIGURE 3.6 All-or-none learning curves that rise at different times, when grouped, yield a gradual, S-shaped curve, which is not representative of any individual's learning.

averages across the same individuals. Notice that each individual curve is of an abrupt, all-or-nothing nature. But the grouped curve is gradual and S-shaped. If interest is in the individual functions, it is folly to group the data.

So why not just plot individual functions in all cases? Unfortunately, this will not always work. Commonly there are confounded variables *within* individuals. Control is exerted across individuals. For example, an experimenter might vary the order of presenting certain stimuli so that each individual gets a different order. The effects of order of presentation of the stimuli would thereby be controlled in the experiment as a whole, but not within any given individual. This method of designing experiments would have to be stopped if only individual functions were used. Furthermore, the variability within individuals

Key Ideas Box 3.6: Limitations on the use of group averages in psychology

Averaging is done on the assumption that deviations around the average may be regarded as errors of measurement. If individual data are of interest, much information will be lost by averaging. Individual functions may differ markedly from group functions derived from the same data (see Figure 3.6).

In order to use individual functions, we must have all variables controlled *within* individuals, and control must be of a high degree so that error of measurement will not obscure true relationships. The psychological method known as *experimental analysis of behavior* takes such an approach.

Sometimes a theory may dictate that we study effects of a variable that are weak relative to the effects of randomly varying factors. This makes the use of individual data virtually impossible. Experimental analysts of behavior are likely to argue in such cases that the theory is less important than studying the influences of the more powerful variables.

There appears to be no way to deal with variables that have irreversible effects (such as brain damage) within the framework of individual functions. Analysis of such functions requires repeated introduction and removal of the variable. This is impossible if irreversible effects are produced.

There is no simple solution to the problem of whether to use individual or group curves. It is most important to realize the difficulties that exist. Thereafter, we can only use the best tools available for a given task.

Logistics Box 3.1: The cumulative recorder

A *cumulative recorder* is a device that bypasses the usual methods of describing data (see Figure 3.7). Responses are recorded directly by having them cause a pen to move across paper that is unrolling at a constant speed. If no responses are made, the pen stands still and a line is drawn parallel to the edge of the paper. The faster the rate of responding, the closer the line comes to being perpendicular to the edge of the paper.

Cumulative records are customarily presented unaltered, with a *legend* to help the eye judge the rate of responding associated with a given angle of line. No descriptive numbers are given to summarize the results. The record is a rather direct reflection of the behavior. However, cumulative plotting has a tendency to conceal variability. It is not uncommon today to supplement the cumulative record with histograms representing the distribution of *interresponse times* (time from one response to another). These provide a more fine-grained analysis.

would have to be controlled-out through meticulous use of technology.

The plotting of individual functions is thoroughly compatible with the methods of experimental analysis of behavior used by operant conditioners. These methods will be discussed in the next chapter. Here it is enough to say that they involve meticulous experimental control within individual organisms. Technological control is exerted to the point where the reality of behavioral changes can be seen without the aid of statistics.

Another facet of the experimental analysis of behavior is its atheoretical bent. Its founder, B. F. Skinner, has spoken ill of theoretically oriented approaches to psychology. If an experimenter wants to test a theory, it is not always easy to do so without using grouped data. Suppose, for example, there are important variables that make individual functions appear somewhat erratic. By averaging across individuals it is often possible, despite variations, to show reliable effects of the independent variable of interest. Thus, group averaging is more important to psychologists with a theoretical bent. Experimental analysts of behavior can take the point of view that theory is less important than establishing functional relationships between important variables. If some other variable is powerful enough to obscure the effects of the variable of theoretical interest, perhaps we should focus attention on the more influential variable and forget about the theory.

The argument of the experimental analysts is a compelling one. It

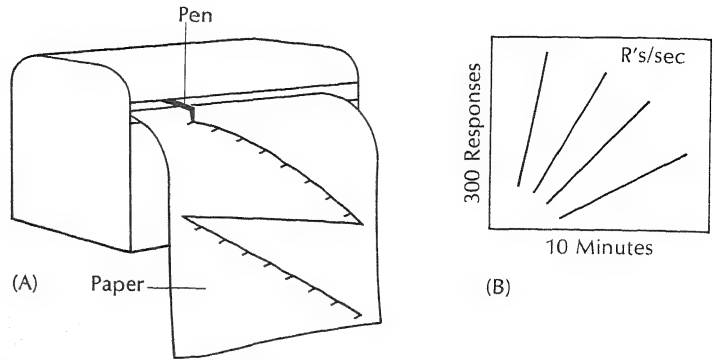



FIGURE 3.7 A cumulative recorder (A). The pen moves a step sideways each time a response occurs. The paper unrolls at a constant speed. The result is that the rate of the behavior is proportional to the angle of the line made by the pen. Reinforcements are indicated by slash marks created by a jiggle of the pen. The legend (B) helps the eye judge the relationship between angle and rate. The paper must be rotated 90 degrees to correspond with the legend.

seems enough to lead one to avoid group averaging whenever possible. However, there are times when we have no choice. For example, we sometimes work with irreversible variables. This means that the variable, once introduced, cannot be taken away. An example would be surgically induced damage to the brain. Brain tissue does not grow back. It is necessary to put in and take out a variable repeatedly in order to work within the framework of the experimental analysis of behavior. So here we *must* use group averages. And sometimes our theory may be important enough that we want to work with subtle variables in spite of the presence of more potent variables.

The most important thing is to realize the difficulties inherent in group curves, and to know that we cannot generalize from individual to group curves. Some psychologists refuse to use any but individual functions. Other psychologists use group averages and ignore the problems inherent in their approach. Perhaps most psychologists are aware of the problems and do the best they can with the tools available to them.

Questions

1. What are Cartesian coordinates?
2. On which axis do we plot the values of the independent variable? The dependent variable?
3. What is the slope of a line?
4. What factors determine a straight line on a graph?
5. What is a histogram?
6. What are positive and negative skew?

7. What is a normal distribution? When is it likely to occur?
8. Define the arithmetic mean, geometric mean, and median, and tell what they measure.
9. What is the standard deviation, and what does it measure?
10. Why can't the sum of the deviations from the mean be used as a measure of dispersion?
11. Give an example of a correction of data and tell why corrections are used.
12. What are linear transformations, and why are they used?
13. What are logarithmic transformations, and why are they used?
14. What are z and Z_i scores, and why are they used?
15. What are the following:
    a. scatter diagram
    b. regression line
    c. regression equation
    d. Pearson product-moment correlation coefficient
    e. Spearman's ρ
16. Distinguish between correlation and causation.
17. Why are psychologists troubled about using group curves?
18. Why do scientists use averages?
19. How does a cumulative recorder work?

4 THE RELIABILITY OF DATA

A scientist's contribution to his field is most often measured by his *findings.* Sometimes scientists contribute in other ways, perhaps by creating a theory; but most often for scientists, as much as for the gold prospectors of yore, findings are what counts. A prospector who first found gold had to worry a bit about the reliability of the find. A naive prospector might go for fool's gold and run into town whooping and hollering about this newfound wealth. Such a novice would come away feeling like a fool. Then, some were taken in by a deceiver who planted gold in order to inflate the value of some land. Or the find might be a very small one, one that was not worth even a few whoops and hollers. Prospectors and scientists, different though they may be, share a common interest in the reliability of their findings.

Scientists are even more dependent on the reliability or repeatability of their findings than are prospectors. Provided the prospectors know enough to tell fool's gold from the real thing, they have at least *something* when they make a find, though it may be worth only a little. But a scientific finding is not paydirt at all if it lacks reliability. In science, a finding is only taken to be fact if it is reliable.

A scientific finding is reliable if it can be repeated (technically we say "replicated") by the discoverer, and also by other experimenters

who have learned the discoverer's method. Reliability signifies the repeatability of findings. But findings will be repeatable if controlling variables have been identified correctly. So the quest for reliability is also a quest for identifying the truly effective controlling variables. If an experimenter's independent variable is effective under the defined conditions of the experiment, the findings will be repeatable. Then the problem of reliability will have been solved.

Two phases in an experiment are critical planning points for reliability. The place to start is before the experiment gets under way. Many things can be done to make it increasingly likely that the results of the experiment will be reliable. The second place is at the end of the experiment. Here, the results are already in, but the experimenter has to assess the reliability of the results. Let us discuss these two facets of reliability.

Maximizing the chances of getting reliable data

A good way to think of strengthening the reliability of data is to consider how to improve on the reliability of ordinary observations and statements. Figure 4.1 summarizes the steps to reliability. Ordinary observations and statements are often surprisingly unreliable. Partly this is due to their vagueness. Hence, *clarity* will improve reliability. In particular, clear definition of terms is important. There is a whole school of philosophy dedicated to the careful analysis of ordinary language. It turns out that controversies, both in philosophy and in everyday life, often dissolve when we analyze and clarify our terms.

To use a down-to-earth example, say that a wife and husband, Jane and Sam, are arguing over the breakfast table. After a period of ill-directed discussion, Sam says, "All right, let's get down to brass tacks! What are you *really* mad about?" Jane replies: "You were hostile to me all last week!" Sam: "I was *not* hostile!" Jane: "Yes you *were!*" As most of you know, this sort of discussion can go on for some time. Jane and Sam might even begin to argue over what *makes* him hostile. "You've been hostile ever since I had a good time dancing with Jim Swave at the New Year's Eve party!"

Usually, we do not think of such human interactions in terms of functional relationships. But, in fact, that is what the discussion is about. Jane is saying that a given set of variables give rise to a psychological state (hostility) in Sam. They both agree that the hostility is not good. Sam does not even want to admit that it is there and feels that he would lose ground by acknowledging it. If it is a functional relationship and if they both agree that the dependent variable (hostile behavior) should be eliminated, why don't they set about manipulating independent variables instead of getting into a furor? One important

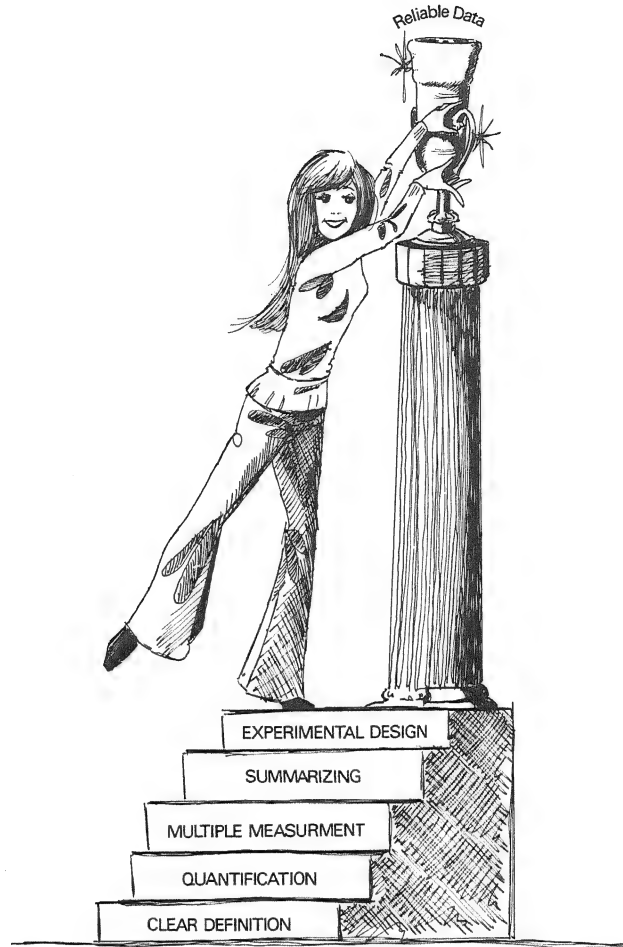



FIGURE 4.1 Steps to the attainment of reliable data.

reason is that they cannot agree that the relationship exists. They cannot so much as agree that the dependent variable occurred.

What happens if they clarify definitions itemizing the hostile behavior? Then Jane must say to herself, "What do I mean by 'hostile'?" The answer takes the form of an enumeration of the activities that fell into the class "hostile": Sam watched TV instead of talking at meals during the past week; he made fun in public of her plumpness; ate lunch out, blaming her cooking; shouted at her over the bridge table; and was "not himself" in other incidents. Now it is more likely that agreement can be reached on the presence or absence of the events, and there is a far greater prospect of predicting and controlling their occurrence. In essence, a shift has taken place, away from a definition of human

activities in terms of inner events and toward clearer, more concrete definition of these activities.

OBJECTIVE CRITERIA AND PROCEDURES

Operational definition

Scientists work toward clarity largely by using *operational definitions.* A term has been operationally defined when we have specified the method of measuring it. We define the thing in terms of its method of measurement. This may seem odd. But operational definitions came into use when various terms in physics proved to be meaningless until their method of measurement was specified. For example, Einstein had trouble with the concept of simultaneity until he used operational definitions.

Here are some examples of operational definitions of terms used in experimental psychology. "Hunger" might be defined as the number of hours elapsed since the last meal, or it might be treated as the number of grams of a given food eaten when it was freely available ("ad lib" feeding). Then again, we might define it as the amount of shock tolerated in order to gain access to a food source. Many other operational definitions could be offered. Since ordinary terms are so vague, they usually end up having many different operational meanings. "Pain" might be defined as the number given when the sufferer is asked to place his pain on a scale with 1 indicating barely detectable pain and 10 indicating intolerable pain. It might also be defined as the amount of increase in blood pressure following the measured introduction of an intense stimulus. It could be defined as the amount the sufferer is willing to increase the painful stimulus before refusing to go further.

In each of these cases we have stipulated a measure and have greatly increased the likelihood that reliable results would occur. An essential stage in preparing an experiment is finding good operational definitions for the concepts we want to study.

Quantification

In many cases operational definitions are already quantified. In these cases the experimenter has taken two steps in the direction of improving reliability. For quantification further helps to achieve the desired goal. Many times we are very accurate about whether an event occurs or not, but are far off base in estimating how often it occurs or how much of it there is. Hence, we may be accurate at one level—the level of saying whether or not it occurs—but inaccurate and unreliable at the level of judging its quantity.

Edwards and Edwards (1970) were interested in the behavior of human fetuses, and used as one measure the "kicking" of the fetus, as

measured by the mother. Mothers, as you know, can feel the movements of the fetus after a certain period of pregnancy. Edwards and Edwards (1970) had mothers and gynecologists estimate how many times the fetus "kicked" each day. Upon measurement, it turned out that fetuses kick many, many times more often than either mothers or gynecologists believed.

In general, then, quantification will improve reliability over and above the improvement obtained with mere clarity of definition.

Multiple measurement

Someone once asked Nobel prize-winner George von Bekesy how he made sure his findings were reliable. He answered that he did so by measuring everything at least four different ways. When we take several different measures of the same thing, we are using the method of multiple measurement or multiple, converging operations. This is different from measuring the same thing several times by the same method. Several different methods of measurement, and therefore several different operations, are used.

Multiple measurement should be used far more frequently than it is. It is a good deal more work, but the end product is worth it. The scientific giants whose efforts provided the foundation of our field often used this method. Go back and have a look at the article by Helmholtz (1852) in which he gives the first measurement of the velocity of the nerve impulse. This was something that virtually everyone, including Helmholtz's own teacher, Johannes Mueller, said could not be done. Yet Helmholtz persisted against the odds. He used equipment seemingly far too primitive for the task, but he found several different ways of measuring the desired velocity, and he took many measurements with each of his methods. Even a finding so hard to fit into the ideas of the time was able to stand up when founded on such masterfully reliable methodology.

Summarizing

We have already discussed methods of describing and summarizing data, but they require mentioning again because they play a role in strengthening reliability.

Devices that permit us to summarize a large body of data in a few numbers can also improve the reliability of judgments. An observer who looks at a large mass of data may be impressed with one part of the data at one time and with a different part at another time. On one day Dr. Fairweather looks at his classroom full of students and thinks they are dullards; on another day, depending on who happens to catch his attention, he may find them very bright. Similarly, different observers may notice different aspects of a large mass of data. For reliability of judgments, summarizing devices are necessary unless the number of

Key Ideas Box 4.1: Procedures which help to assure that data will be reliable

The following factors are helpful in producing reliable data:

1. CLEAR DEFINITION

When definitions are at the level of ordinary, casual description, disagreements about the facts commonly occur. Making definitions more clear by moving away from abstractness and toward concreteness is likely to increase reliability. It is valuable to stick close to observable events in making definitions. Operational definitions, in which a thing is defined by stipulating the method of measuring it, are especially useful.

2. QUANTIFICATION

Even when definitions are clear, additional reliability may be obtained through quantification procedures such as counting or scaling data. The more exacting the measurement, the more reliable the data.

3. MULTIPLE MEASUREMENT

By measuring the same thing in more than one way reliability can be increased. This procedure has an advantage over measuring the event many times in the same way because it will tend to reveal discrepancies that result from variations in measurement method.

4. SUMMARIZING

Since unreliability can result from looking selectively at a given aspect of a large mass of data, descriptive statistics, which summarize the raw data, can improve reliability. Early experimenters sometimes tend to "pick out the best data," and inadvertently end up with unrepeatable findings. Use of averages and measures of variability around the average helps to prevent this.

5. EXPERIMENTAL DESIGN

It is necessary to identify controlling variables in an experiment if we are to specify the conditions under which the experiment can be repeated. The presence of confounded variables will make this an impossible task. Since a major function of experimental design is to minimize the risk of introducing uncontrolled variables, these designs make a major contribution to the assurance of data reliability.

observations is very small or the observations are virtually identical to each other. The importance of descriptive statistics, such as the mean, median, standard deviation, and range, in improving reliability should not be underestimated. In a number of concrete instances their contribution has been very important to scientific advancement.

For example, not long ago it was the custom for physiologists and physiological psychologists dealing with the evoked electric potentials of the brain[1] to select and present "representative" recordings of the potentials for publication. This sort of technique can lead to a very distorted view of the typical evoked potential. Today there is widespread use of such devices as the "computer of average transients," which summarizes the result of hundreds of evoked potentials. The transition from selection of so-called representative individual data to the use of averages that take all of the data into account enhances reliability. Of course, this is true far beyond the area of evoked potential recording. It is, in fact, common in any research area for early experimenters to rely on selection of data on an intuitive basis until summarizing procedures come to be introduced. The influence of the intuitive procedure during this early phase on the reliability of published data cannot be assessed. It could be catastrophic.

If we use summarizing statistics *across* individuals, the statistics may not represent the individuals at all well. We get better representation of the data with summarizing statistics provided we do not thereby conceal data of interest. We average when we intend to treat deviations around that average as "error." If we are not willing to treat such deviations as error, it may be better to avoid summary statistics for the data in question.

DEALING WITH EXTRANEOUS VARIABLES

Suppose you are setting out to answer a question experimentally. You have to analyze your question, and find measurement operations for the concepts of interest. You can then go looking for a reliable functional relationship. What you really want is to find independent variables that work. Your major adversary in this quest is the uncontrolled variable. If uncontrolled variables interfere, you won't know what, if any, potency independent variables have. You have two major weapons to use against the uncontrolled variable. They are *technology* and *chance.* If you want to get rid of an uncontrolled variable, you may seek the technology to do so. If you lack the technology, you may arrange things so that the uncontrolled variable will be evenly represented in the various treatments. Chance will work for you because

[1]An evoked electrical potential is a change in the voltage of a part of the brain following sensory stimulation of the subject.

random processes tend to distribute themselves evenly in the long run. But you have to be sure that chance gets an opportunity to operate. You have to look out for systematic errors. In so doing, you are engaging in experimental design. Experimental design is discussed in Chapter 5, but some underlying principles will be discussed here.

What can you do about an uncontrolled variable? First, you might consider *allowing it to influence your results.* If so, you are deliberately confounding. At times, this is not a bad idea, but it limits your understanding of controlling variables. On the other hand, since real-life variables are complex, it may give you a more authentic variable. We discussed deliberate confounding in more detail earlier, and so will stop with it here.

The most common way of coping with a confounded variable is to try to *hold it constant.* You can do this through technology or statistically. Technologically you might *reduce it to zero,* eliminate it. This is holding it constant at zero. Or you might *hold it constant at above zero.* For example, if you want to find out whether a certain breed of dogs can discriminate color, you cannot reduce brightness to zero. The dogs couldn't see anything. But you could make each color equally bright. (Technically this is hard to do. This is because apparent brightness depends on the visual system of the perceiver. You cannot assume that what is bright for you is also bright for a dog.)

Often, you cannot hold a variable perfectly constant, but only *approximately* so. This goes for reducing it to zero, too. But although physicists cannot make a perfect vacuum, nevertheless they still use vacuums. Approximation is often used in science.

Now let us suppose you want to hold a variable constant statistically. You can do this either by randomization or rationally. For example, suppose I am worried that the time of day when subjects participate in an experiment is important. I cannot make the day stand still. I lack the technology. But I can see to it that all the different times of day are equally represented for each of the treatments. If I determine randomly which treatment each arriving subject will get, any time of day will eventually occur equally often across treatments.

But maybe you do not want to trust to chance. Chance will rarely turn up things in some systematic order. But sometimes it will. Sometimes poker players get a royal flush. In the long run, this is not worth worrying about. But in the short run, it may be smart to distrust chance. So you may want to use a rational procedure for averaging out the variable of concern. Going back to concern over the time of day, you could always conduct your experiment at the same time each day or deliberately see to it that each time of day was equally represented in each treatment. You could just schedule subjects so that each treatment was given equally often at each running time.

Randomization is good if you have a large number of things to

randomize. Randomization may even out variables you have not thought about. But randomization can only be expected to hold variables constant in the long run. If you have a small number of observations, rational procedures may well be better.

A final way to deal with uncontrolled variables is to *manipulate* them. Students in one of my laboratory courses were doing an experiment on the two-point threshold. They were trying to find the smallest detectable separation between two points on the skin. If the two points are close enough together, they are perceived as one. Move them apart further and further, and they will eventually be perceived as two. Several students got widely scattered, unreliable results. They felt that variations in the pressure of the points might have caused the nonuniformity. What were they to do? They could have used any of the

Key Ideas Box 4.2: Dealing with extraneous variables

There are two basic ways to control for the effects of extraneous variables. These are technology and chance. If you have the technology, you may simply control the influence directly. If not, you may have to rely on distributing the influences evenly over control and treatment conditions through statistical means.

There are several specific ways of dealing with such variables. *Deliberate confounding* means allowing them to influence the results. This limits knowledge of controlling variables, but allows you to deal with the kinds of complex clusters of variables that occur in nature. If you do not wish to tolerate such confounding, you can *hold the variable constant,* either at zero (eliminate it) or at some positive value. You can also hold it *approximately constant.* Approximations often serve well enough to get the scientific task done.

Variables may be held constant through *randomization.* If, for example, subjects are assigned randomly to treatment conditions, results will even out in the long run. Randomization cannot be relied upon to hold a variable constant when numbers of observations are small. A second option is to use a *rational procedure* where the opportunities to influence results on the part of the extraneous variable are deliberately arranged so as to occur equally often in all conditions.

A final way to deal with such variables is through *manipulation.* The extraneous variable can be treated as a second independent variable and manipulated systematically to see whether it has an influence on outcomes and just what that influence is.

methods for dealing with uncontrolled variables. But one group decided to rig up a device to vary the pressure systematically. This would show whether this variable was important. If so, the two-point threshold should be different at different pressures. Manipulation is a good way to deal with uncontrolled variables. It not only gives you control, but also shows limitations on when the functional relationship holds.

Assessing reliability

We will soon discuss how to design experiments. But before designing something, it is a good idea to ask yourself what you intend to do with it. Experiments give outcomes that will have to be evaluated for reliability. How do we assess reliability? I will begin by telling you a story.

> Sally Howe stepped from her car and walked briskly into the house. She shivered a little as she closed the door behind her. The leaves had just begun to fall, and it had suddenly become a little chilly. Besides, she felt tense because Bill was away, and there had been a number of housebreakings in town over the past few weeks. "Why can't Bill get a job that allows him to stay home once in a while?" she thought peevishly. Then she abruptly turned on herself. "Stop it, Sally!" she thought, "You know that's the perfect job for Bill, and besides you shouldn't be so darned dependent on him!"
>
> She set her lips tight in a gesture of resolve and walked quickly upstairs to the bedroom. She closed the bedroom door and locked it behind her. Sally was tired, and hurriedly prepared herself for bed, got under the covers, and turned out the light. Her fatigue overcame her tension, and she soon drifted into that world halfway between dreams and waking.
>
> But suddenly she sat up with a start. There was a creaking sound downstairs! She felt her heart pound fast. "My God! Is there someone down there?" she half-whispered. She listened intently for a full minute. "There it is again!" she thought.
>
> "But no. It's just this old house creaking the way it always does when the weather gets cold," she said to herself. "You've simply *got* to quit acting like a silly child."
>
> She lay back down, closed her eyes firmly, and tried hard to go to sleep. Then there was a sound like someone on the stairs. She bolted up and grabbed the telephone by the bed. She started to dial. But as her fingers spun the dial she began to feel that the frightening sound blended in with the other creakings of their old house. She put down the phone. Her eyes began to moisten. "I don't want to act hysterical," she thought, "it may just be the boards bending in the cold . . . but *what if someone is down there!*"

Sally's predicament is like that of an experimenter. Both of them

want to know whether certain happenings are due to something important. In both cases, they have to worry whether the data might be due to uncontrolled fluctuations of uninteresting factors. They have to ask themselves: "Am I getting an important message, or is this just due to noise in the background?"

Sally doesn't want to call the police because she wishes to avoid seeming silly, hysterical, and dependent. Technically, we would say she wants to avoid a *type I error.* She makes a type I error if she sounds the alarm when no one is really there. This is like an experimenter publishing an article saying that an independent variable worked when it did not. Both are type I errors; both are "false alarms."

But there is another kind of error. What if someone is really there and Sally fails to call the police? This is like the experimenter who fails to detect it when an independent variable works. This kind of error is called a *type II error.* It is a "miss."

A major point brought out by the story of Sally is that both she and the experimenter are trying to detect a signal in the presence of a noisy background. Sally wants to know whether certain sounds in the house are different enough from creaking of the background to merit being considered a signal of danger. The experimenter likewise wants to know whether results gotten with a certain experimental treatment differ enough from what happens without the treatment to merit reporting that the variable is effective.

Basically, we assess reliability of data by deciding whether our independent variable had an effect. And we decide whether it had an effect by comparing it to some noise level, some level of fluctuating values that occurs without the independent variable. We must discriminate between a noisy background alone and a noisy background when there is a signal present.

We distinguish the combination of signal + noise from noise alone by asking ourselves how likely it is for the observed event to occur if there is just noise. Sally heard a certain level of noise. Some of the sounds were enough to make her worry that there might be a prowler. The noises seemed a bit too loud to be just part of the background noise, even in her creaky old house. We can imagine a sudden crash and the sound of a lamp falling over and Sally saying, "That's it, there *is* a prowler!" Theoretically a lamp could fall over without a prowler, but *that* is so unlikely that Sally will act on the assumption that someone is down there. Experimenters do the same thing. If the results in their treatment condition are markedly different from what occurs, or could reasonably be expected to occur, without the treatment, they take it for granted that the variable had an effect.

Technically, you can think of the judgments of Sally and the experimenter as based on a ratio. For Sally it is the following:

$$\frac{\text{Observed creaking}}{\text{Creaking without prowler}}$$

If the observed creaking is about the same as creaking without the prowler, the ratio is about 1. If the observed creaking is a lot more than the usual creaking, the ratio will be large. Implicitly, Sally has a certain cut-off point in her head. If the ratio exceeds that cut-off value, she will call the police.

For the experimenters, the ratio is, roughly,

$$\frac{\text{Observation}}{\text{Error estimate}}$$

They look at the observed difference between treatment and control conditions, and contrast it with differences that occur without any treatment.

A general way of saying what Sally and the experimenter are both doing is that they want to tell the ratio

$$\frac{\text{Noise}}{\text{Noise}}$$

from the ratio

$$\frac{\text{Signal} + \text{Noise}}{\text{Noise}}$$

If what they observe is a case of

$$\frac{\text{Signal} + \text{Noise}}{\text{Noise}}$$

the ratio should be larger than if it is

$$\frac{\text{Noise}}{\text{Noise}}$$

If the ratio is large enough, they decide there is a signal.

How, in practice, do experimenters find their ratios and decide whether to say their findings are reliable? The first way is by inspection.

JUDGING RELIABILITY BY INSPECTION

Sometimes the ratio of observation to noise is so great that we can just look at the data and decide that there must be a signal. This happens if there is a *whopper variable.* A whopper variable is one that produces striking effects, clearly beyond the noise level. For example, we might find that there is no overlap at all between the results of an experimen-

tal and a control group. Say, everybody given a new teaching method did better than anybody given the old method. If no overlap occurred despite many observations, or even if a little overlap occurred, we could decide by inspection that the variable had an effect.

We can sometimes by inspection tell that data are reliable even when we lack a whopper variable. This happens when the noise level is very low. Even a small signal may be recognizable on a background of little or no noise. Sally would have an easier time judging whether a prowler was there if she lived in a very quiet house.

Mere luck is rarely enough to give us low noise levels. We need meticulous experimental control for this. By bringing behavior under careful control prior to introducing treatment conditions, we make effects of the independent variable obvious by inspection.

There is a whole school of psychologists who emphasize use of careful experimental control to the point where reliability can be judged by inspection. We call their approach the *experimental analysis of behavior.* This approach has been espoused by B. F. Skinner, and people interested in operant conditioning use it most often.

With an experimental analysis of behavior, the experimenter establishes a *baseline* of behavior to which performance is contrasted when the independent variable is put in. The baseline is usually stable and relatively free of noise. These experimenters often have to take a very long time getting the desired baseline behavior. They commonly work with small numbers of experimental subjects, and each subject receives both the experimental and control treatment. (This is called a *within-subject design.*) They can afford to take a great deal of time and work

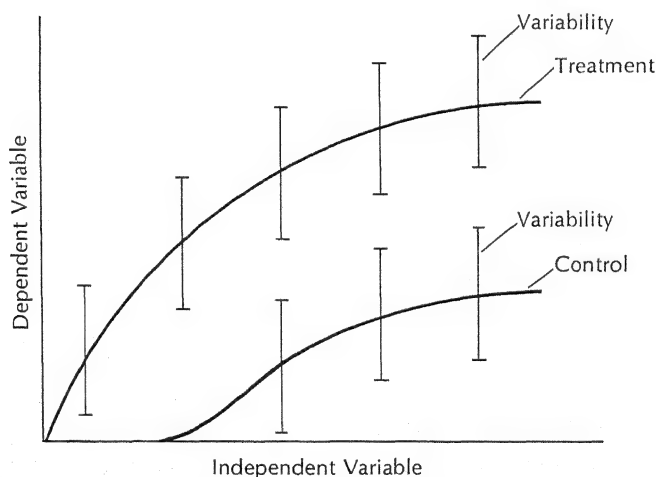



FIGURE 4.2 Idealized graph of the effects of a "whopper variable." The variable is so powerful that there is no overlap between the data with it and the data without it. The two curves are far enough apart so that they do not overlap despite high variability.

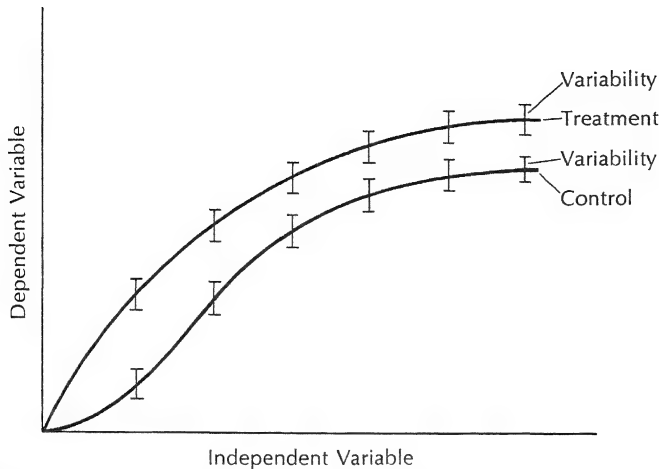



FIGURE 4.3 Idealized graph of effects produced under high experimental control. The two curves are close together, but the variability is so slight that the two sets of data do not overlap.

with few subjects because the effects are made so clear. The high degree of experimental control compensates for the relatively reduced number of subjects.

STATISTICAL JUDGMENTS OF RELIABILITY

The eye is a relatively insensitive instrument for judging reliability. The influence of many effective independent variables would be missed if we did not compensate for this insensitivity. Reducing the noise level by exerting high degrees of experimental control helps. Another method is the use of statistics. Psychologists seem more prone to sharpen the eye with statistics than with experimental control. There is a controversy over which method is best, but we will not deal with that here.

When something gets too subtle for the eye to discriminate, scientists usually turn to instruments that enhance its power. Thus they turned in the past to the microscope and the telescope. The idea behind statistics is to improve on the ability to discriminate the effects of experimental treatments. Many experiments have outcomes not obviously reliable by inspection, but reliable nevertheless. The reliability of such outcomes can be assessed with statistics.

Risks of rejecting the null hypothesis

What is it that we use to decide that the independent variable had an influence? Think of Sally. She had to judge how likely were the sounds she heard, assuming there was no prowler. Certain sounds, like the

simplicity, let's say she has recorded 1000 sounds. When a given sound occurred, she could (theoretically) compare it to the 1000. If a sound as great or greater than that had occurred 212 times, she would probably say, "See, this kind of thing happens all the time. There's no good reason to think there is a prowler." But what if that great a sound has only occurred once before in the normal course of events? She might then say, "The chances are one in a thousand that this is just background noise. I'm not taking such chances with my life. I'll assume there is a prowler."

What we are doing here is comparing the observed event to what can be expected if there is only background noise. In experiments, this would mean comparing our finding to what can be expected from random fluctuations and error of measurement. We temporarily assume that there is only error of measurement, and find out how often our observation could be expected to occur if our assumption were true. This is the same as saying we assume the independent variable had no effect. We then find out how likely our outcome is on that assumption. The assumption that the independent variable had no effect is called the *null hypothesis.*

With the experimental analysis of behavior, we also find out what would happen if the independent variable did not work. We do so by running subjects for a long time without the variable. The baseline behavior is behavior without the independent variable. With statistics we try to save ourselves some work by doing at least part of this mathematically. The mathematician finds a model of the situation and finds out by computation how often the observed outcome would take place without an effective independent variable.

How can a mathematician do this? Let's take a concrete example. Suppose you want to know whether your pet dog, Gordon, can discriminate color. You get a bunch of colored cards and, as a control for brightness, some grey cards. You order Gordon to sit, then place one card on your right and one on your left. Holding one of his favorite treats in hand, you say "Here, Gordon." Gordon comes running over and jumps up on you, ignoring your cards. But you "No, no!" him until he catches on that he is to nuzzle cards. You've decided to start him out on red and green. If he picks red, he is correct and you give him a treat and pet him. If not, you say "No, no!" and make him go sit. You keep changing the cards from right to left at random from trial to trial. Periodically you use the grey cards and deprive him of the treat no matter what he does. That way he should learn that you're not interested in his picking, say, the darker card.

After a few days, Gordon runs to the red card 7 out of 10 times. You ask yourself, "Is he doing better than chance?" What is the likelihood of getting 7 out of 10 correct by chance? Assuming that the red cue has no influence on Gordon, how often will he do that well? A very

Key Ideas Box 4.4: Null hypothesis and errors of inference

The notion that the independent variable had no effect is called the *null hypothesis.* We determine the reliability of the independent variable's effect by contrasting it to outcomes expected if the null hypothesis were true. In the experimental analysis of behavior this is done by running subjects many, many times without the independent variable and contrasting the results with those obtained under the influence of the independent variable. In contrast, the statistical approach is to estimate the results that would occur in the long run without an effect of the independent variable. Experimental and statistical analyses both operate by contrasting null outcomes to those actually observed with the independent variable.

Basically, both procedures entail looking at a ratio of the effect observed *with* the independent variable to the effect *without* it. The variation found without it may be considered noise. The effect with it is either noise alone or signal plus noise. We want to decide whether there is a signal. If there is not, the ratio should be close to 1. If it is larger than that, we have evidence for a signal. How large does it have to be? We decide this arbitrarily.

The *alpha level* of probability is a function of the size of the ratio that we accept for rejection of the null hypothesis. It is typically set at 0.05, which means that one time in twenty we expect to reject the null hypothesis when it is true, to give a false alarm. False alarms are called *type I errors.* The alpha level is actually the probability of a type I error. Another type of error is to miss a signal when it is really there. A miss is called a *type II error.* The probability of a type II error is called *beta.* It is a set of numbers rather than a single number.

laborious way to find out would be to eliminate the redness of the card and run him a thousand trials or so to find the proportion of times he performed as well as 7 out of 10 correct. But it would be easier to use a handy model. Without the red cue, there would be two equally likely choices. Anything with two equally likely outcomes could be used as a model. Why not flip a coin? Let heads indicate a correct response and tails mean an incorrect response. It would be easier to flip coins than to struggle with Gordon a thousand times. But have you ever tried flipping a coin a thousand times? It's boring.

The easiest thing is to use a mathematical model. In the eighteenth century Blaise Pascal worked out the mathematics of such situations

for some gambling friends. We call the correct model the *binomial distribution.* Table B in the Appendix shows how likely it is that Gordon would get 7 correct in 10 tries if he were not relying on the red cue. The probability is 0.172.[2] That means that you can expect to get this result about 17 percent of the time, even if Gordon has learned nothing about color. This is hardly a solid basis for deciding he has color vision.

You have to decide what probability that results are coincidental you are willing to treat as low enough to rule out chance alone. Logically, this decision is arbitrary. But in practice, we generally accept only probabilities of 0.05 or lower. The probability we set for rejecting the null hypothesis is called the *alpha level.* The alpha level tells us the proportion of time we can expect to be wrong in rejecting the null hypothesis. At the 0.05 level we can expect to reject the null hypothesis incorrectly 5 percent of the time, or once in every 20 experiments. Rejecting the null hypothesis incorrectly, remember, is called a type I error.

To be wrong so often in thinking the independent variable had an effect seems intolerable to many people. They like to set their probability level somewhat lower. Maybe you would prefer once in a hundred experiments (0.01 level) or once in a thousand (0.001 level). But when you lower your alpha level, you increase the risk of failing to detect an independent variable that works. This is called a type II error. The probability of a type II error is called *beta.* It is, incidentally, not a single number, but a set of numbers. Its value depends on properties of the population that we normally do not know.

Power and deciding how many subjects to use in an experiment

If we fail to detect a real effect of our independent variable, we have committed a type II error. It is important to have experimental designs and methods of evaluating their outcomes that make it likely that we will detect effects when they are there. If our method is unlikely to miss effects, we say it is of high *power.* "Power," then, is the technical term for *sensitivity to real effects of an independent variable.* Statisticians define power as 1 minus beta, where beta is the probability of missing an effect of the independent variable (type II error). The combined probabilities of detecting an existing effect and of failing to detect it must equal 1. If the probability of a type II error is, say, 0.50, the probability of detecting a real effect (the power of the method used) must be 1 minus 0.50 = 0.50.

[2]Actually, the situation is a bit more complex than we are pretending. Gordon is repeatedly getting 10-trial opportunities. We have calculated as though he got only one. But, for simplicity, let's leave out the mathematical details here.

Experimental tactics differ in their power. Several things can be varied to influence power. The most obvious of these is the alpha level, or probability of a type I error. Compare the influence of setting alpha at 0.001 and at 0.10. At 0.001 you are less likely to make a false alarm, to wrongly reject the null hypothesis. But this conservative criterion will also make it more likely for you to miss effects when they are there. Going back to Sally, if she insists on very strong evidence of a prowler, she will be unlikely to embarrass herself by calling the police when no one is downstairs. On the other hand, the risk of allowing a prowler free reign goes up when the demand for evidence is so high. Power could be increased by choosing the alpha level of 0.10. In Sally's case this would be like her calling the police when the evidence was not very strong. She would be unlikely to fail to detect a prowler. On the other hand, she would be highly likely to make a false alarm. Thus, we can conclude that, although power can be increased by lowering the alpha level, this method of increasing power is not very desirable because it increases the rate of false alarms or type I errors.

Fortunately, there are other, more desirable ways to increase power. Anything that increases the size of the ratio of observed effect/noise increases power. For example, if the size of the signal is large, the ratio will be larger. We can deliberately choose to study highly influential, whopper variables. But this, too, has its drawbacks because it tends to limit the range of things we can study. A better approach is to concentrate on making the bottom half of the ratio, the noise level, smaller. One way to do this is to increase the number of observations we make in the experiment *(N).* This can be done by including more subjects in the experiment or by taking more measurements on the subjects we already have. The larger the number of measurements, the more stable will be our estimate of the actual population values. Increased stability lowers noise.

One of the most desirable ways to increase power is use of technology. As we manipulate the experimental environment, we can do so in ways that decrease noise level (error of measurement). Anything that minimizes the influence of the small extraneous variables that contribute to noise levels will improve power. Thus, exerting greater experimental control is a highly desirable way to improve power.

Reducing the noise level is exactly what experimental analysts of behavior do. Otherwise, they would have methods of very low power, since they depend on inspection to judge reliability. Statistical tests vary in power, but it would be hard to imagine one less powerful than mere inspection. High degrees of experimental control are usually needed to compensate for it. Experimental analysts of behavior also make very large numbers of observations. Both of these factors save them from low power.

STATISTICAL POWER ANALYSIS: A NEGLECTED TOPIC IN EXPERIMENTAL DESIGN. There would appear to be little point in doing an experiment that has small likelihood of permitting detection of effects *that are really there.* Yet, behavioral scientists commonly do so by designing experiments of low power. Cohen (1962) did a survey of research in abnormal–social psychology and found that the experiments typically had only enough power to detect effects of very large magnitude at the 0.05 alpha level. For the cases in which real effects were medium or small, Cohen (1962) estimated that the respective probabilities of failing to detect them were 0.52 and 0.82. The situation would be still worse for alpha values smaller than 0.05.

Behavioral scientists often choose sample sizes according to criteria unrelated to test power. It is common to use some arbitrary size such as the number of subjects readily available, the number of subjects used by previous experimenters who did related research, or the number of subjects the particular experimenter has used in the past. Such procedures are nonrational and dangerous. Cohen (1969) has provided tables of statistical power analysis that make it possible to arrive somewhat rationally at decisions regarding sample size, given a prior decision concerning alpha and a rough estimate of the anticipated effect size.

Often the sample sizes needed to have a reasonable degree of power are far greater than those we see in published articles. We need to be more careful about determining sample size. We should also learn a good lesson from the experimental analysis of behavior. Finding out how to control baseline behavior is very important, whether we prefer the experimental analysis of behavior or not.

It is not easy to decide rationally on a sample size. Dixon and Massey (1957) suggested the use of sequential designs. With a sequential design, we determine the sample size while conducting the experiment. We start out with a small sample, and having completed a block of the experiment decide (1) to reject the null hypothesis, (2) that the null hypothesis cannot be rejected, or (3) to increase the sample size.

Sometimes the first block of observations permits rejection of the null hypothesis. If so, it is all to the good. We have achieved our goal with great economy of effort. Keep in mind that rejection of the null hypothesis on the basis of a few observations is more impressive with respect to reliability than rejection on the basis of many.[3] This is because the test is of low power when there are few observations. Only very strong effects will be detected this way.

At other times, the overlap between experimental and control treat-

[3]Unfortunately, decisions made on the basis of small samples may be unsafe with respect to *generality.* Direct or systematic replication by conducting a coherent series of experiments will remedy this. These are discussed at the end of the chapter.

Key Ideas Box 4.5: Power and size of sample

The power of a test is its sensitivity to signals, to real effects of the independent variable. Since beta is the probability of missing a real effect, power is 1 − beta. Power is influenced by the size of the real effect (is it a whopper variable?), sample size, the degree of noise in the experiment, and the arbitrary alpha level. There is evidence that many experiments done in psychology are of low power. This could be raised by increasing the degree of experimental control or increasing sample size. We have no control over the size of the real effect (except to choose whopper variables for investigation), and lowering the alpha level would increase type I errors.

How do we decide what sample size to use? Most investigators simply choose a size used by other investigators working in the same area. This is not necessarily safe. Cohen (1969) shows a way to estimate the sample size needed to give adequate power. His method entails several estimations that seem unsafe. Doing experiments sequentially is another alternative. This means conducting a block of the experiment, then deciding whether to reject the null hypothesis, to add another block of subjects to increase power, or to redesign the experiment more judiciously.

ments may be so great at the end of the first block (or *n* blocks) that a virtual miracle would be needed to reject the null hypothesis, even in the long run. At such times it is wise to quit and rethink your experiment.

Then there are those times when the results are leaning in the direction of permitting you to reject the null hypothesis, but they have not reached the stipulated alpha level. Here you can go on to make another block of observations.

Statistics and reliability: The F *ratio as an example*

We have said that judgments of reliability are based on a ratio. It is some sort of ratio of observation to noise. If the ratio is large, if it exceeds a certain criterion, we decide the data are reliable. The task of statistics is to make the calculation of this ratio more exact than it is when we judge by merely looking at the data.

The *F* ratio is a good example. The *F* comes from the name of the great statistician R. A. Fisher, who worked it out. He based the ratio on variances. Remember, the variance is a measure of deviation from or dispersion around a mean. It is the mean of the squared deviations

Fisher used a ratio with variances between groups divided by variances within groups. Looking at it closely, you can see that this is just another ratio of observation to the experiment's noise level.

Suppose someone has discovered a drug purporting to make people more intelligent. You have it in the form of pills—call them "smart pills." You want to test whether they work. Since they might be effective in some doses but not in others, you decide to use two widely different doses and a placebo control. Six women and six men agree to participate in the experiment. You randomly assign four subjects to each group, with the constraint that the sexes must be equally represented in each group. You find an excellent test to use as a dependent variable. Call it the "Schlepper Test of Mental Acuity." Results are as follows:

Placebo Subject Score	Low Dosage Subject Score	High Dosage Subject Score
Sam 102	Cecilia 133	Kelly 140
Mary 111	April 124	Josh 141
June 107	Mike 120	Holly 152
George 120	Bill 125	Louie 143
Means: $\bar{X}_P = 110.0$	$\bar{X}_L = 125.5$	$\bar{X}_H = 144.0$
Grand Mean ($\bar{T}$) = 126.5		

Now let's take a look at Louie. His score is 143.0. If the smart pills worked, and it looks as though they might have, his score is due to two different things. One of them is error of measurement, noise. The other is a true effect of the pills. We can view his score as a deviation from the Grand Mean ($\bar{T}$). The Grand Mean is simply the mean of all twelve scores, ignoring which group they came from. Louie's deviation from $\bar{T}$ is $143.0 - 126.5 = 17.5$. If his score of 143 were purely a true measure, we could say that all of this deviation was due to an actual effect of the pills. But we know that any real observation is a mixture of the true measure and error of measurement. So this total deviation of 17.5 could be *partitioned* into a part that is due to true effect and a part that is noise.

How can we get an estimate of what is noise? A good estimate of noise would be the amount of deviation from the mean with the influence of the pills taken out. But notice that *within* each subgroup that influence of the pills is either absent or held constant. Why not take as our estimate of noise some measure of the deviations of scores within a subgroup/from the subgroup mean? Louie's deviation from his subgroup mean is $143.0 - 144.0 = -1$. This gives us a basis for estimating how much of his 17.5-point deviation from the Grand Mean is due to noise.

Maybe now is a good time to remind you that we cannot use the

simple deviations with their signs as a measure of dispersion. These deviations around the mean will always sum to zero. So we will square them and find the mean of their squares, using the *variance* as our measure of deviation. The F ratio turns out to be a ratio of variances.

Now that we have a basis for a noise estimate, we can put something in the bottom (denominator) of the ratio. But what shall we put in the top (numerator)? We need an estimate of the deviations (variance) due to the independent variable. We are treating $\overline{T}$ as the estimate of the true over-all mean. The obvious thing is to use the mean of each subgroup as the estimate of the true mean for a given drug condition. There are three subgroup means, and they can be treated as scores deviating around $\overline{T}$.

We took out the influence of the independent variable and calculated variance within groups to get an estimate of noise. Conversely, to get an estimate of the effect of the independent variable we must take out the noise, the variance within groups. Taking the mean for each group does just that. The deviations *within* groups are of individual scores around the group means. Now we calculate a variance *between* groups. For each group, we take the group mean, subtract it from $\overline{T}$, and square the result. This gives us a squared group deviation from $\overline{T}$. That squared deviation is due to differences between groups, with differences within groups taken out. The resulting number can be applied to each subject in the group, to represent this subject's contribution to the variance between groups. Thus, the squared "between-groups" deviation for a given group must be multiplied by the number of subjects in that group.

To get the total sum of squares between groups you must add the between-groups sums of squares for each group. So, for your experiment on smart pills, you would figure a between-groups sum of squares for the Placebo Group, for the Low Dosage Group, and for the High Dosage Group. You would then add them all up. Thus

$n_{Pl}(\overline{T} - \text{Mean}_{\text{Placebo}})^2 + n_{LD}(\overline{T} - \text{Mean}_{\text{Low Dosage}})^2 +$
$n_{HD}(\overline{T} - \text{Mean}_{\text{High Dosage}})^2 =$
Total sum of squares between groups.

Now, keep in mind that the variance is an *average* or *mean* of squared deviations. The number obtained at this point is a sum of squared deviations. To get a mean, you sum, then divide by something. To get the variances for the F ratio, you divide by *degrees of freedom*. We will discuss this concept next.

Descriptive versus inferential statistics

Sometimes we use a statistic to describe a limited sample of subjects we have observed. At other times, we use a statistic as an estimate of a true measure not really observed. This true measure is called a *parameter*. It is a characteristic of a population rather than a sample.

Key Ideas Box 4.6: The *F* ratio

The *F* ratio is a ratio of a figure derived from observed differences between groups over an estimated noise level. It helps determine whether an observed difference between groups is signal plus noise or noise alone. It is a ratio of variances. Variance within groups provides an estimate of noise. Variance between groups provides an estimate of signal (if there is a signal) plus noise. Thus, variance *between* groups divided by variance *within* groups is a measure of how far the results obtained under the influence of the independent variable deviated from those obtained by noise alone.

Thus the mean number of drinks per day of a sample of alcoholics is a statistic. It provides an estimate of the actual mean number of drinks per day of *all* alcoholics, which is a parameter.

The very same statistic—mean, variance, and the like—may serve two distinct functions. One function is to describe observed data. The other function is to provide an estimate of a parameter.

Of course, we want the best possible estimate of a parameter. In some cases the way of calculating a descriptive statistic is the same as the way of calculating the same statistic used as an estimate. For example, the mean of a sample is the best estimate of the parametric mean. The formula for the mean stays the same whether we use it to describe a sample or to make inferences about a parameter.

Sometimes, though, the formula must change when we change from descriptive to inferential use of the statistic. This is the case with the variance and standard deviation. The best estimate of the parametric variance is not the descriptive variance. A slight change in the formula is necessary. Instead of dividing by the sample size, we must divide by the number of *degrees of freedom* (abbreviated df). Then we have the best estimate of the parameter.

How do we know how many degrees of freedom there are? The number of degrees of freedom used to divide a given sum of squares[4] is equal to the number of scores with *independent* information on which the sum of squares is based. Put another way, degrees of freedom equal the number of scores less the number of estimated parameters used in calculating the sum of squares.

But why do we say that the number of scores with independent information is smaller than the total number of scores? Take this

[4]The term "sum of squares" is a short form commonly used for "the sum of the squared deviations from the mean." It is the top part of the formula for the variance. The number of degrees of freedom is the bottom part.

example. We have three scores, 10, 12, and 8. The mean of these equals 10. We estimate the population mean to be 10 ($10 + 12 + 8 \div 3 = 10$). Now we calculate a sum of squares by subtracting each score from 10, squaring the difference, and summing the three resulting squares. Ten minus 10 equals zero, 12 minus 10 equals 2. Is our last score free to vary? Not if the means equals 10 and sample size is three. The only number it can be is 8. Hence the last score does not give us independent information. The number of degrees of freedom in this case is equal to $3 - 1 = 2$.

The variance between groups is based on one estimated population mean, $\overline{T}$ (the Grand Mean), and the means of each of the groups, treated as scores. In our "smart pill" example, there were three groups. The number of degrees of freedom for the variance between groups is therefore again $3 - 1 = 2$. In general, the number of degrees of freedom for the variance *between* groups is equal to the number of groups minus one. Calculation of variance *within* groups is based on deviations around the means of each individual group. In the "smart pill" experiment, there were three groups of four subjects each. Three means were used as estimates. Thus, instead of taking the total number of subjects (12) as the divisor for the sum of squares of deviations from the respective group means, we must subtract one for each estimated mean. So we subtract one per group. Hence the number of degrees of freedom within groups is equal to the total number of subjects minus the number of groups. In our example this is $12 - 3 = 9$.

We started out by saying that Louie's deviation of 17.5 from $\overline{T}$ could be broken down into two parts. These were the deviation between groups (signal) and the deviation within groups (noise). This says that the total deviation is equal to the sum of deviations from these two component sources. The total variance is equal to the sum of the variance between groups plus the variance within groups. How many degrees of freedom are there for the total sum of squares? The deviation values on which it is based are obtained by subtracting each individual's score from $\overline{T}$. Hence the number of degrees of freedom equals the total number of scores minus one for the estimated parameter ($\overline{T}$).

The F ratio is the mean square (MS) between groups over the mean square within groups.

$$F = \frac{MS_{between}}{MS_{within}}$$

Each mean square is simply the sum of squares divided by its appropriate number of degrees of freedom.

F ratios for various alpha levels may be found in the appendix, Table E. The value depends on the numbers of degrees of freedom for the top

Key Ideas Box 4.7: Descriptive versus inferential statistics and degrees of freedom

When a statistic is used to describe a limited sample, its formula sometimes differs from that of the same statistic when it is used to estimate values in a population from values in a sample. The first is called a descriptive statistic; the second is called an inferential statistic. The formulas for inferential statistics provide *best estimates* of population values.

The formula for the mean as a descriptive statistic is the same as the formula for the best estimate. But the formula for the variance differs slightly, depending on whether it is descriptive or inferential. Whereas with the descriptive statistic the squared deviations from the mean are divided by the number of observations (N), they are divided by the number of *degrees of freedom* (df) when the variance is used inferentially. Degrees of freedom equal the number of scores less the number of estimated parameters used in calculating the numerator of the variance. Population means must be estimated to calculate the inferential variance, so the number of means used in the calculation has to be subtracted from N to get degrees of freedom.

degrees of freedom and will say more in the next chapter. The specific details of calculating an F ratio are given in Chapter 5. When you get to that chapter, why not analyze the data from the "smart pill" experiment?

JUDGING RELIABILITY BY REPLICATION

In the long run we look at the results of more than one experiment to evaluate data reliability. In fact, some published studies actually consist of many experiments. If several organisms are used and each receives identical treatment, as is often the case, the data for each organism can be considered a separate repetition of the basic experimental unit. In the Russian journals it is not uncommon to list data from each subject as a separate experiment. This makes a great deal of sense, provided the experiment is complete and not confounded within subjects. In some experimental designs, where control is across subjects, this kind of procedure would make less sense because the basic unit of information is the average of the group. For any given subject in such an experiment, important variables are confounded and therefore repetition of the experiment requires that the same or, more often, a

different group be run through the experiment again. For obvious reasons, the two types of repetition (or *replication*) just described are called *intersubject replication* and *intergroup replication* (Sidman, 1960). With intersubject replication the experiment done on one subject is repeated on other subjects and the results are compared. With intergroup replication the experiment done with one group of subjects is repeated with another group and the resulting averages compared.

Replication does not always consist merely in the exact repetition of an experiment. Such repetition is called *direct replication.* It is often a better strategy to start a new experiment, which yields new information but also shows whether the previous results can be replicated.

For example, I once ran an experiment that included taking measures of the ability of albino rats to perform on visual discrimination tasks with one eye when the task was learned with only the other eye open (Sheridan, 1965). Such transfer from one eye to the other is called interocular transfer. When I moved to a new university, I wanted to replicate the experiment. One strategy would have been to move the original apparatus to the new location, obtain rats from the same colony, run the rats myself, and in general conduct the new experiment as nearly as possible in the same way as the original one. Instead, I had a new apparatus built that was similar but not identical to the original one, a different strain of albino rats was used, and a colleague ran most of the rats. If the results were like those of the original study, the experiment would not only show that the original results could be replicated, but also that the original results were general despite variations in rat strain, experimenter, and the other changed factors. If, on the other hand, the results were not repeated, it would be necessary to go in search of the relevant variables.

Such a search can entail much work, but it will tell a great deal about the relevant variables in the original study. The worst that can happen, then, is that variables previously thought to be irrelevant will reveal their influence. And that is not bad at all. On the other hand, obtaining results essentially similar to those of the original experiment would mean that the effort entailed in a direct replication had been spared. The results of the experiment described above were in fact virtually identical to the original ones (Sheridan & Shrout, 1965).

When we conduct a new experiment that will also show the replicability of an earlier one, the procedure is called *systematic replication* (Sidman, 1960). It is a case of systematic replication when we use the results of one experiment to create a baseline for a new experiment. For example, an early discovery that organisms rewarded only occasionally for responses are very persistent in responding when rewards are withdrawn has provided a baseline which is often used to study the effects of a wide variety of variables. Each new study is a systematic replication.

Key Ideas Box 4.8: Replication

In establishing reliability of results, nothing is superior to repeating an experiment (replication). Replication may be *intersubject*, in which a controlled experiment done on one subject is repeated with others, or *intergroup*, in which a controlled experiment done on one group of subjects is repeated with another group. Replication may be *direct*, which means that the same experiment is repeated in all its details. On the other hand it may be *systematic*. This means that a new experiment contains conditions that will tell whether the first experiment had reliable and general results. For example, a finding from the first experiment may provide a baseline for a second experiment. Systematic replications provide evidence for both reliability and generality. However, failure in systematic replication may require considerable search for the responsible variables, since several things have changed at once.

Study Questions

1. Explain each of the following and tell how they help improve reliability:
   a. clarity of definition
   b. operational definition
   c. quantification
   d. multiple measurement
   e. summarizing
2. What are the two major weapons against uncontrolled variables?
3. List the main ways in which experimenters control variables.
4. Explain type I and type II errors.
5. Under what two conditions can reliability be judged by inspection?
6. What method is used in the "experimental analysis of behavior" to evaluate reliability?
7. What is the null hypothesis?
8. What is the relationship between judgments of reliability made by experimental analysis of behavior and by statistical methods?
9. What is meant by alpha level?
10. What is beta?
11. What is power of an experimental method?
12. What are the main methods of selecting size of samples?
13. What is a sequential design?

14. What is an *F* ratio? How does it relate to the concepts of signal and noise?
15. What is a parameter?
16. What are degrees of freedom?
17. Distinguish between descriptive and inferential statistics.
18. Explain the following:
    a. intersubject versus intergroup replication
    b. direct versus systematic replication

Exercise

1. Get a recent issue of *The Journal of the Experimental Analysis of Behavior, The Journal of Experimental Psychology,* and *The Journal of Personality and Social Psychology.* Note the devices used in each of these journals to establish reliability. Do you see patterns in the use of baseline or statistics? How often do you notice the use of direct and systematic replication?

5 DESIGNING EXPERIMENTS

When a finding has been established reliably, the variables responsible for the finding have been isolated. It is by knowing which variables control an outcome that we are able to reproduce the outcome at will. And it is the ability to reproduce an outcome that is the essence of reliability.

How can we find out which variables control an experimental outcome? The obvious way is to exert experimental control—to be sure that the proper control groups have been included in our experiment. Yet it is not always easy to avoid overlooking a control that should have been run. Even the best experimenters have trouble keeping in mind all possible controls. Furthermore, an experimenter wants to maximize the sensitivity of the experiment and to minimize its cost. It is not easy to achieve both those goals. Fortunately, this burden can be lightened by keeping in mind standard models or *experimental designs.*

Experimental design tells the experimenter how to assign subjects to the various experimental and control conditions.[1] It provides an outline showing how to arrange the experiment with proper controls. It also provides a handy language system for telling others what has been

[1]Note that this is different from sampling procedures. You may be tempted to confuse the two. Sampling is a matter of selecting subjects for study. Experimental design tells how to assign the selected subjects to treatments. We discussed sampling procedures in the section on the generality of data.

done. An experiment might be very cumbersome to design and to describe if done without consideration for typical designs. But it might be relatively easy to accomplish both design and description if the experimenter simply decided on a "2 × 2 × 2 factorial design" (see "Factorial Designs" later in this chapter). Hence, it is worthwhile for anyone who expects to read, design, or describe experiments to learn certain widely used model designs. Here we will give a very broad outline of major classes of design.

Completely randomized designs with two groups

A completely randomized design has two characteristics. Each treatment (experimental or control) is given independently to a different sample of subjects. Samples of subjects are drawn independently and are assigned randomly to experimental and control conditions. What is meant by "samples of subjects are drawn independently"? In some types of design assignment of a given subject to one of the conditions *determines* that another particular subject (for example, a twin) be assigned to the opposite condition. In the completely randomized design, placing a given subject in one condition of the experiment has *no influence* on the placement of any of the other subjects. Figure 5.1 illustrates this type of design. Let's consider an example.

Suppose you wonder whether people can learn to see the world correctly when they have distorting prisms placed over their eyes. Experimenters have been curious about this kind of question for a long time. Originally they thought it might tell us why we see the world upright despite the fact that the image on the retina is upside down. You might also want to know the conditions under which relearning occurs. There is evidence that it makes a difference whether subjects are active or passive (see, for example, Held, 1965). So suppose you decide to compare the effects of active or passive movement on adaptation to distorting prisms.

All subjects could be placed in a visual environment with distorting prisms over their eyes. One group could be composed of subjects allowed to move around on their own initiative. The other group could be moved about passively in a wheelchair. Some measure of adaptation to the prism-induced distortion could be taken. The experimental and control group adaptation-performances could then be compared.

There may be some advantage in going over this type of experiment in some detail, step by step. First, you obtain a group of subjects by one of the various means available to experimenters. Typically they draft students enrolled in introductory psychology courses, hire subjects through an advertisement, or something of this sort. Theoretical problems of finding subjects have to do with *sampling procedures*, and these are distinct from experimental design.

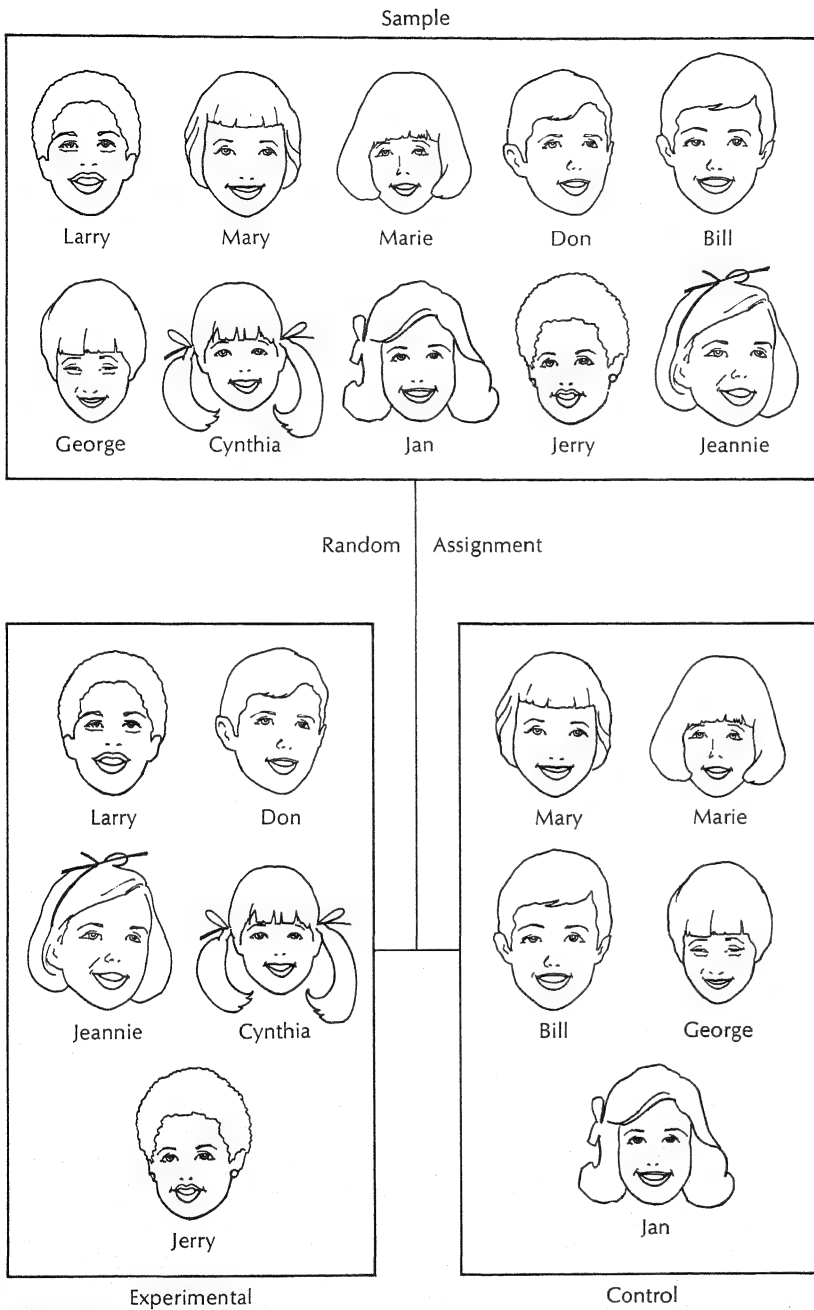



FIGURE 5.1 Completely randomized design with two groups. Subjects are randomly assigned to conditions.

You must decide whether a given subject will be assigned to the active condition or to the passive wheelchair condition. There are a number of ways you can do this. Haphazard methods tend, inadvertently, to produce systematic patterns. One effective way would be to compile a list of the subjects' names, placed in any order at all, and to assign subjects to conditions by way of random numbers.

RANDOMIZING BY WAY OF A TABLE OF RANDOM NUMBERS

The best way to randomize subjects, treatments, or what-have-you is to use a table of random numbers. You will find one in the Appendix, Table A. Start with a list of the subjects in any order. Assign each subject a random number by entering the table at a haphazard starting place, and taking the numbers in order after that. You must make up a rule for deciding which numbers belong in which group. You might say that even-numbered subjects go in the experimental group and odd-numbered subjects go in the control group. Any other property of the numbers that would give you a two-way split would be all right.

Experimenters use tables of random numbers often. The tables are useful with types of experimental design other than those using independent subjects. You should get used to working with such tables.

The following is a list of imaginary subjects. Next to each subject is a random number I found by entering the table in a haphazard way and moving straight down the column thereafter.

Cynthia	3	Experimental
George	8	Control
Larry	9	Experimental
Mary	2	Control
Jerry	1	Experimental
Don	5	Experimental
Jeannie	7	Experimental
Jan	2	Control thereafter
Marie	7	
Bill	3	

The third column shows the subject's assigned group. Notice that after five subjects had been assigned to the experimental group by chance, the rest had to go in the control group. This constraint was imposed by the wish to put equal numbers of subjects in each group. The process was random with that constraint. We could tolerate unequal numbers of subjects in each group, but it would make the analysis difficult. Also, with small samples serious inequalities might occur. One of the groups might be too skimpy to provide a good estimate of the real effect for that condition.

The design ends up looking as follows (see Figure 5.1):

Experimental Group	Control Group
Cynthia	George
Larry	Mary
Jerry	Jan
Don	Marie
Jeannie	Bill

Purpose of randomization

Experiments are designed to detect effects of the independent variables and to detect only those effects. It might seem that the best way to control for other variables would be to identify them and eliminate their effects. But to identify all variables that might conceivably influence an experimental outcome would not be easy. By randomizing, we arrange an experiment so that extraneous factors tend to be equally represented in experimental and control groups. Coin tossing will tend in the long run to yield equal numbers of heads and tails. Random assignment to conditions in an experiment will tend to produce equal representation of variables requiring control.

ANALYZING DATA FROM A COMPLETELY RANDOMIZED DESIGN WITH TWO GROUPS

After conducting the experiment, you will have to assess the reliability of the findings. We have described how this is done in general terms. But a specific kind of experimental design calls for a specific type of analysis. Most of the time, there will be enough error of measurement to require statistical analysis. Detailed rationales for such analyses are found in statistics texts. Here, I will tell you which types of analysis are

Key Ideas Box 5.1: Randomization

We have random assignment when every possible sample of subjects has an equal and independent chance of being assigned to a given treatment. The purpose of randomization is to equalize the influence of noise-producing factors over the various conditions. Randomization works in the long run. It may not work for small samples.

There are various methods of randomization. Experimenters commonly use tables of random numbers to guide them in randomizing. It is not enough to use a haphazard method, since this commonly leads to systematic patterns. A method of using random numbers is explained in the text.

appropriate. Simple descriptions of how to compute the statistics are in logistics boxes. None of this is a substitute for a good statistics course.

Various methods of analysis could be used with this kind of design. Most often you would use a t test for independent subjects. "Cookbook" methods of computing this and other statistics are given in logistics boxes, in this case Logistics Box 5.1. The data are imaginary and may vary from box to box. If possible, you *should* use the t because it is the uniformly most powerful test. This means that it is the most sensitive instrument for detecting real effects. Certain assumptions were made in devising the t test. The variances of the experimental and control groups were assumed to be equal. This means that error of measurement is the same whether the independent variable is present or not. The data were also assumed to come from a population having a Gaussian distribution. Recent studies have indicated that the t test remains accurate despite sizeable violations of these assumptions. It can be used unless the violations of assumptions are fairly large.

There is a test almost as powerful as the t test that is not based on these assumptions. And it is very easy to compute. It is called the *Mann-Whitney* U *test* (see Logistics Box 5.2). You will find it very useful for rapid analysis of data.

Incidentally, statistical tables are sometimes based on one-tailed and at other times on two-tailed probabilities. A one-tailed alpha merely takes into account one possible direction of difference between the treatment conditions (either greater or less but not both). Two-tailed probabilities take into account both possibilities. You will rarely want other than the two-tailed figure, since readers tend to be suspicious of the more lenient one-tailed test even when it is logically justified. In general, two-tailed values can be obtained from a one-tailed table simply by doubling the associated alpha level. For example, a one-tailed p of 0.01 yields a two-tailed p of 0.02.

The appropriateness of a statistical test (for example t versus U) among other things, depends on the type of measurement of the dependent variable. With the t test we assume that the data are measured on at least an interval scale. The U test requires only that the data be on an ordinal scale.

Completely randomized designs with more than two groups

So far the discussion has centered on experiments with only two groups of subjects, an experimental and a control group. The completely randomized design can also be used for a greater number of groups. In fact, it can have n groups of subjects. For example, students could be independently and randomly assigned to three procedures for teaching introductory psychology. One group could receive a standard lecture course. A second group might get an opportunity to learn the same material with a teaching machine. A third group might be given a

Logistics Box 5.1: Calculating *t* for independent means

The *t* ratio equals the difference between the two means divided by their standard error.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{SE_{\bar{X}_1 - \bar{X}_2}}$$

1. Find the difference between the means.

Experimental Scores	Control Scores
2	0
3	1
1	1
2	2
3	2
$\Sigma X_1 = 11$	$\Sigma X_2 = 6$

$$\bar{X}_1 = \frac{\Sigma X_1}{5} = 2.2 \qquad \bar{X}_2 = \frac{\Sigma X_2}{5} = 1.2 \qquad \bar{X}_1 - \bar{X}_2 = 2.2 - 1.2 = 1.0$$

2. Find the standard error.

$$SE_{\bar{X}_1 - \bar{X}_2} = \sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \text{ where } S^2 = \frac{\left(\Sigma X_1^2 - \frac{(\Sigma X_1)^2}{n_1}\right) + \left(\Sigma X_2^2 - \frac{(\Sigma X_2)^2}{n_2}\right)}{n_1 + n_2 - 2}$$

a. Square each score and sum the squares for each group.

4	0
9	1
1	1
4	4
9	4
$\Sigma X_1^2 = 27$	$\Sigma X_2^2 = 10$

b. Square ΣX_1 and ΣX_2 and divide each by the corresponding n.

$$(\Sigma X_1)^2 = 11^2 = 121; \frac{121}{n_1} = 24.2$$

$$(\Sigma X_2)^2 = 6^2 = 36; \frac{36}{n_2} = 7.2$$

c. $S^2 = \frac{(27 - 24.2) + (10 - 7.2)}{5 + 5 - 2} = \frac{5.6}{8} = 0.7$

d. $SE_{\bar{X}_1 - \bar{X}_2} = \sqrt{0.7\left(\frac{1}{5} + \frac{1}{5}\right)} = \sqrt{\frac{1.4}{5}} = \sqrt{0.28} = 0.53$

4. $t = \frac{1.0}{0.53} = 1.89$

5. Degrees of freedom $= n_1 - n_2 - 2 = 8$ at $df = 8$

Table D (Appendix) shows that t at $df = 8$ must be at least as large as 2.306 to correspond to the 0.05 alpha level. Thus t is too small to permit rejection of the null hypothesis.

Logistics Box 5.2: Computing Mann-Whitney *U*

1. List the scores.

Experimental Scores	Control Scores
5	2
5	1
2	0
4	1
3	3
5	
6	

2. Select the group with larger scores and for each score, tally how many scores in the other group are larger than it is. Ties count 0.5.

Scores	No. larger
5	0
5	0
2	1.5
4	0
3	0.5
5	0
6	0

3. U equals the sum of the resulting tallies.

$$U = 2.0$$

4. n_1 = number of cases in the smaller group; n_2 = number of cases in the larger group,

$$n_1 = 5,\ n_2 = 7$$

Table C (Appendix) shows the probability of getting U as small as 2.0 with these n's is 0.01. The null hypothesis can be rejected. Methods for large samples are given by Siegel (1956).

regular text, but simply study it at home. The assignment of subjects to conditions is essentially the same whether two or *n* groups are used.

THREE-GROUP RANDOMIZATION

Again, let's illustrate the assignment procedure with imaginary subjects. We will start with 15 subjects and require that there be 5 subjects in each group. This time we cannot use oddness or evenness of

numbers as a basis for assignment. We need to divide the numbers into three equally likely categories. I chose to use the numbers 1 through 9 in the table of random numbers. The numbers 1, 2, and 3 represented assignment to the lecture method. The numbers 4, 5, and 6 indicated assignment to the teaching machine method. The rest of the subjects would then have to get the home-study method.

First, we arrange a list of the subjects. I used the following list:

Cynthia	Don	Ruth
George	Jeannie	Donna
Larry	Jan	Lilly
Mary	Marie	Sam
Jerry	Mike	Gordon

This list is not randomly arranged. It is in the order in which the names came to my mind. People cannot generate random processes unaided. Going into the table of random numbers I got: 5, 0, 9, 5, 7, 8, 9, 0, 7, 7, 8, 3, 4, 5, 8, 8, 1, 9, 5, 9, 1, 5, 2, 2.

The first number is a 5, so the first subject, Cynthia, goes in the Machine group. The second number is zero. It is not from 1–9, so ignore it. The next number is 9, so the second subject on the list goes into the Home Study group. I continued this process until, at the ninth random digit, all the Home Study slots had been assigned. After that I ignored, besides 0, the numbers 7, 8, and 9. By the digit third from the end, all the Machine slots were assigned. The remaining subjects were assigned to the Lecture method.

To be sure you understand how to do this, it would be good practice to finish up the assignment of the listed subjects and arrange them under the headings: "Lecture," "Machine," and "Home Study."

ANALYSIS OF THE COMPLETELY RANDOMIZED DESIGN WITH MORE THAN TWO GROUPS

The most common statistical method for analyzing data from completely randomized designs with more than two groups is the simple, one-way analysis of variance (see Logistics Box 5.3). This is based on the F statistic we already discussed. You should be aware that the relationship between the earlier discussion of the F statistic and the computational procedures in the logistics boxes may be far from obvious. Formulations designed to explain the logic of statistical procedures are commonly quite different from those most convenient for computation. For example, the top part of the formula for variance (Chapter 4) is

$$\sum_{i=1}^{n} d_i^2 \quad \text{or} \quad \sum_{i=1}^{n} \frac{(X_i - M)^2}{n}$$

but the most convenient form for computing from raw ("X") scores is

$$\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}$$

In effect, this says that

$$\sum_{i=1}^{n} d_i^2 \qquad \text{equals} \qquad \Sigma X_i^2$$

minus a correctional factor,

$$\frac{(\Sigma X_i)^2}{n}$$

The analysis of variance assumes that data are on an interval scale. It also assumes they are selected from a Gaussian distribution of measures. Finally, it assumes that the variances of the different groups are equal. However, here as with the t test, studies indicate that the statistic works well despite sizeable deviations from the assumptions, provided sample sizes in the different groups are equal (see Keppel, 1973). In fact, the t test is merely a special two-group case of the F test.

If data are only on an ordinal scale or if you are concerned about extreme deviations from assumptions, a Kruskall-Wallis one-way analysis of variance is appropriate. For data on a nominal scale (frequencies), a Chi-squared test for k independent samples will do. These two analyses are explained very clearly and simply in Siegel (1956).

Logistics Box 5.3: Analysis of variance for single-factor, completely randomized design

1. Obtain the total sum of squares (SS_T).
   a. Obtain the correction factor (C) by summing all the scores, squaring the results, and dividing by the number of scores (n).

Levels of Factor

A Scores	B Scores	C Scores
1	2	3
2	3	4
1	2	3
0	1	2
2	3	4
$\Sigma X_A = 6$	$\Sigma X_B = 11$	$\Sigma X_C = 16$

$$\Sigma X_i = 6 + 11 + 16 = 33$$

$$(\Sigma X_i)^2 = 1089$$

$$C = \frac{(\Sigma X_i)^2}{n} = \frac{1089}{15} = 72.6$$

b. Find SS_T by squaring each score, summing the squares, and subtracting C.

A Scores	B Scores	C Scores
1	4	9
4	9	16
1	4	9
0	1	4
4	9	16
$\Sigma X_A^2 = 10$	$\Sigma X_B^2 = 27$	$\Sigma X_C^2 = 54$

$$\Sigma X_i^2 = 10 + 27 + 54 = 91$$

$$SS_T = \Sigma X_i^2 - C = 91 - 72.6 = 18.4$$

2. Obtain sum of squares between groups (SS_b).

$$SS_b = \frac{(\Sigma X_A)^2}{n_A} + \frac{(\Sigma X_B)^2}{n_B} + \frac{(\Sigma X_C)^2}{n_C} - C$$

$$= \left(\frac{6^2}{5} + \frac{11^2}{5} + \frac{16^2}{5}\right) - 72.6 = 10.0$$

3. Obtain sum of squares within (SS_w).

$$SS_W = SS_T - SS_b = 18.4 - 10 = 8.4$$

4. Find ms_b by dividing SS_b by df_b.

$$df_b = \text{number of groups} - 1 = 3 - 1 = 2$$

$$\frac{SS_b}{df_b} = \frac{10}{2} = 5$$

5. Find ms_w by dividing SS_w by df_w.

$$df_w = \text{number of scores minus number of groups} = 15 - 3 = 12$$

$$ms_W = \frac{SS_W}{df_w} = \frac{8.4}{12} = 0.7$$

6. Find F by dividing ms_b by ms_w.

$$F = \frac{ms_b}{ms_W} = \frac{5}{0.7} = 7.14$$

7. Read Table E (Appendix) for F corresponding to alpha $= 0.05$ at 2 and 12 df. The $F = 3.89$, so the null hypothesis can be rejected.
8. It is customary to summarize results of analysis in a table like this:

Source	df	SS	MS	F
Total	14	18.4	—	—
Between	2	10.0	5.0	7.14
Within	12	8.4	0.7	—

Designs with matched subjects and two groups

We assess reliability by looking at a ratio of

$$\frac{\text{signal} + \text{noise}}{\text{noise}}$$

If this ratio is near one, we assume that there is just

$$\frac{\text{noise}}{\text{noise}}$$

If it is sufficiently greater than unity, we decide there is a signal. If there is a signal, this means the independent variable had an effect, and if it had an effect it will have one in the future as well, all else being equal.

Anything we do to reduce the bottom part of the ratio makes us better able to detect effects. Technically we would say that decreases in the estimate of error increase the power of our design or test. In practice, a major source of noise is the tendency for individual subjects to differ from each other. We have very different backgrounds; we come from very different genetic pools. Some people have stomachs many times the size of other people's stomachs. This is bound to influence their eating behavior. Some people get agitated when given drugs that stimulate most of us. Other people are sedated by stimulants. The many such individual differences are likely to inflate the noise factor.

Designs with matched subjects are used in an attempt to reduce the influence of individual differences. Subjects are not assigned to groups randomly. They are matched on some factor that is positively correlated with the dependent variable. Figure 5.2 illustrates this type of design.

As an example of the procedure of matching, suppose we are interested in the effects of caffeine on memory. We might match our subjects prior to the experiment according to their grade-point average (GPA). Our assumption would be that GPA and memory are positively correlated. We would then give one member of each matched pair the experimental treatment and the other the control treatment. An experimenter interested in the effects of marijuana on chess playing might be fortunate enough to get sets of identical twins closely matched in chess-playing performance. One twin could be given the measured dose of marijuana, while the other received some sort of placebo. A third common example of matching is the use of littermate controls in animal experiments. For example, Held and Hein (1963) were interested in measuring the effects of active vs. passive rearing on later visuomotor performance. Pairs of kitten littermates were selected; one of each pair was randomly assigned to the active condition and the other to the passive. (Chapter 9 tells the outcome of their study).

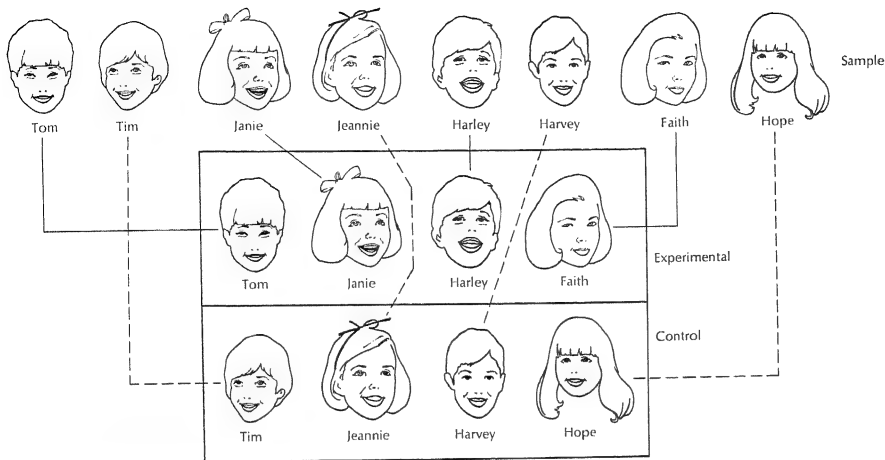



FIGURE 5.2 Matched subjects design with two groups. One each from matched pair is assigned to experimental and control groups.

Notice that randomization is used even with this type of design. Assignment to the experimental or the control treatment is random, with the constraint that no more than one of each matched pair can be in a given treatment condition.

SETTING UP A DESIGN WITH TWO GROUPS OF MATCHED SUBJECTS

Suppose we set out to conduct the experiment on the effects of marijuana on chess playing. We obtain a group of identical twins evenly matched for ability in chess. They are:

Tim	Tom
Janie	Jeannie
Harvey	Harley
Faith	Hope

For each pair, we want one of them assigned to the experimental group, which receives a measured dose of marijuana just before playing. The other goes into a control condition with a placebo. First, let's decide on a rule for deciding which of each pair gets the marijuana. Let's say that we will select four random numbers. If the number is even, the subject listed first gets the experimental treatment. If it is odd, the subject listed second gets the treatment.

Entering the table of random numbers, I got 5, 6, 3, 6. Five is odd, so Tom gets the treatment. Six is even, so Janie gets the treatment. By the same rule, Harley and Faith get the treatment (see Figure 5.2).

Precautions in using designs with matched subjects

Since matching reduces noise, it might seem that every experiment should be done with matching. But there are certain drawbacks of matching. For one thing, when you give up pure randomization, you may inadvertently introduce systematic variations in other relevant variables of which you are unaware. Such confounding could be devastating. A second problem is that you must be sure that the variable used for matching has a *positive* correlation with the dependent variable. If the correlation happens to be negative instead of positive, the result will be an inflation of the noise level, instead of a reduction. The reasons for this are a bit complex to treat here, but the effect must be a matter of concern nevertheless.

Yoked controls

One method of getting a very precise matching of subjects is to use *yoked controls.* With this technique, control subjects receive the same stimulation as experimental subjects. But the experimental subject has control over the stimulation and the control subject does not.

An example of yoked controls comes from a study of the development of vision in active versus passive kittens (Held & Hein, 1963). Students of perception have long felt that there is some special relationship between action and vision. It was a very long time ago when Bishop Berkeley (1685–1753) proposed that vision develops as a result of moving in the environment. Held and Hein wanted to raise kittens with equal visual experience, but with one group of subjects actively moving in the visual world, and the other group passively moved. They devised a *kitten carousel* that yoked pairs of kittens together. One kitten had its feet on the ground, so its movements were translated into changes in its visual world. The other kitten moved in the opposite direction whenever the active kitten moved. The amount of visual exposure was the same for each kitten.

ANALYSIS OF DESIGNS WITH MATCHED SUBJECTS AND TWO GROUPS

The statistic used most often for designs with two groups of matched subjects is the *t* test for correlated means (see Logistics Box 5.4). This test is much like the *t* test already discussed, except that it corrects the estimate of error for the reduction of noise due to matching.

For ordinal data a simple, but not very powerful, test is the *Sign Test* (see Logistics Box 5.5). With the Sign Test, you simply put down a plus for every case in which measures of experimental subjects exceeded those of their matched subject (or vice versa—the direction doesn't matter). The result is a list of plusses and minuses. The binomial

Logistics Box 5.4: Computation of *t* for matched (or same) subjects

The *t* ratio equals $\dfrac{\bar{D}}{SE_{\bar{D}}}$

1. Find $\bar{D}$, the mean of differences between matched subjects. List pairs of scores, subtract one from the other, and find the mean of the resulting differences.

Pairs	Experimental	Control	D_i
1.	2	0	+2
2.	3	1	+2
3.	1	1	0
4.	2	1	+1
5.	3	2	+1
			$\Sigma D_i = 6$

$$\bar{D} = \frac{\Sigma D_i}{N} = \frac{6}{5} = 1.2$$

2. Find $SE_{\bar{D}}$.
   a. Square each difference and sum the squares.

$$\begin{array}{r} 4 \\ 4 \\ 0 \\ 1 \\ 1 \\ \hline \Sigma D_i^2 = 10 \end{array}$$

   b. Compute standard deviation of difference.

$$S_D = \sqrt{\frac{\Sigma D_i^2}{n} - \left(\frac{\Sigma D_i}{n}\right)^2} = \sqrt{\frac{10}{5} - \left(\frac{6}{5}\right)^2}$$

$$= \sqrt{0.56} = 0.75$$

   c. $SE_{\bar{D}} = \dfrac{S_D}{\sqrt{n-1}} = \dfrac{0.75}{2} = 0.38$

3. $t = \dfrac{1.2}{0.38} = 3.16$

4. Degrees of freedom $= n - 1 = 5 - 1 = 4$. The tabulated value for alpha $= 0.05$ at 4 df $= 2.776$, so the null hypothesis can be rejected.

Logistics Box 5.5: Computation of sign test

1. List the pairs of scores and make a + next to those pairs where the experimental value is greater than control, a minus when it is less, and a zero for ties.

Pairs	Control	Experimental	Sign
1.	2	3	+
2.	1	2	+
3.	2	3	+
4.	0	2	+
5.	1	3	+
6.	2	1	−
7.	0	3	+
8.	3	3	0
9.	1	2	+
10.	2	3	+

2. Throwing out tied pairs, determine the proportion of remaining pairs with the less frequent sign (+ or −).

   There is one minus, so $x = 1$
   There are 9 untied pairs, so $n = 9$

3. Look up the probability of $x = 1$ with $n = 9$ in a binomial table (Table B in Appendix).

   $p = 0.02$, one-tailed

4. For most purposes, double the tabulated probability, converting to a two-tailed test.

   $p = 0.04$, two-tailed

5. If p is less than the stipulated alpha (usually 0.05), reject the null hypothesis.

   $p < 0.05$, therefore reject the null hypothesis

distribution (in the Appendix, Table B) can be used to see whether the proportion is significantly deviant from chance.

A more powerful alternative to the t test for correlated means is the *Wilcoxon Matched-Pairs, Signed-Ranks Test.* It is to be used for ordinal data, since it takes into account the ranks as well as the mere direction of the data, it is more powerful than the Sign Test. The method of computing it may be found in Siegel (1956).

Designs with matched subjects and more than two groups

RANDOMIZED BLOCK DESIGN

We can use the matching principle on any number of groups. The *randomized block design* illustrates this. It is also called a "treatments by levels" or "stratified" design. Subjects are matched prior to the experiment on some measure that is positively correlated with the dependent variable. They are then arranged in two or more blocks. Similarity of subjects within blocks is thus greater than similarity of subjects between blocks.

Suppose you are interested in the ability of people to control their heart rate voluntarily. You decide to manipulate the type of feedback they get while trying to learn such control. You have a control group, a group receiving points on a counter for correct control (this is simple biofeedback), a group receiving 1¢ per beat for correct control, and a group receiving 5¢ per beat.

Let's say you have read Luria's *Mind of a Mnemonist* (Luria, 1968). In this book Luria describes his many years of studying a man with seemingly flawless photographic memory, who could control his heart rate markedly by picturing himself running down a street. Exceptional visual imagery seems positively related to voluntary control of heart rate. Variations in the ability to make visual images might well be an important source of variance in your study.

You could use a completely randomized design, but the variable of "imagery" would contribute to your estimate of noise. Why not measure ability to make visual images prior to the experiment, and then put subjects into three blocks with high, medium, and low imagery?

Remember from our discussion of the *F* ratio that the noise estimate, the bottom half of the ratio, is based on deviation scores of subjects from the means of their treatment groups. This is the within-groups variance. By using means within blocks for calculating these deviations, the variance can be reduced. The following example shows how this works.

First let's set up a completely randomized design. I will make up absurdly simple numbers for the heartbeat changes.

TABLE 5.1 Completely Randomized Design

Control	*Counter*	Feedback	
		1¢ per Beat	*5¢ per Beat*
0	1	2	3
0	1	2	3
0	2	3	4
0	2	3	4
0	3	4	5
0	3	4	5
Means = 0	2	3	4

Each score represents the heartbeat change for a given subject. The deviation scores around the means for each group are as follows:

TABLE 5.2 Deviation Scores around Means

Control	*Counter*	Feedback	
		1¢ per Beat	*5¢ per Beat*
0	−1	−1	−1
0	−1	−1	−1
0	0	0	0
0	0	0	0
0	1	1	1
0	1	1	1

A variance within groups could be calculated by squaring each of the deviation scores, summing them, and dividing by the number of degrees of freedom.

Now let's set it up as a randomized blocks design.

TABLE 5.3 Variance within Groups

Imagery	Control	*Counter*	Feedback	
			1¢ per Beat	*5¢ per Beat*
	0	1	2	3
Low	0	1	2	3
Means =	0	1	2	3
	0.	2	3	4
Medium	0	2	3	4
Means =	0	2	3	4
	0	3	4	5
High	0	3	4	5
Means =	0	3	4	5

Now if you calculate the deviations around the means within blocks, they are all zero! See how the variance can be lowered by blocking? A drawback, of course, is that degrees of freedom are reduced. Remember, the degrees of freedom equal the number of observations minus the number of estimated parameters. The estimated parameters are the means. Four means were used with the completely randomized design. Twelve were used for the randomized block design. Hence degrees of freedom will undergo substantial reduction with the randomized block design.

Is it worth tolerating this reduction in degrees of freedom? Normally it is, provided the variable used for blocking the subjects has a reasonably high positive correlation with the dependent variable. The correlation should probably be at least +0.2 (Feldt, 1958).

ANALYZING THE DATA FROM DESIGNS WITH MATCHED SUBJECTS AND MORE THAN TWO GROUPS

In most instances analysis of data from designs with matched samples and more than two levels of the independent variable will be based on analysis of variance. The analysis is different from that used with completely randomized designs. Computational procedures are shown in Logistics Box 5.6.

For data at the ordinal scale level, the *Friedman Two-Way Analysis of Variance* is appropriate. For data at the nominal scale level, the *Cochran* Q *Test* can be used. Both of these tests are explained in Siegel (1956).

Designs with repeated measures on the same subjects

Matching one subject with another can never be perfect, though it can be very good in some cases. The most thoroughgoing extension of matching procedures can be realized by giving the same subject more than one treatment. In such cases, repeated measures are taken on the same subjects. It would be difficult to imagine a more satisfactory "matching" procedure, since subjects are certainly more similar to themselves than to any other subject, even an identical twin.

Just as in the case of the matched-subjects design, the goal of a same-subject design is to reduce error variance by reducing intersubject variability. If variability between subjects can be reduced, the noise level of the experiment will be lowered. Thus the effects of a given independent variable will be more readily identified. The same-subject design is very powerful. A further advantage of this sort of design is that it permits observation of data from individual organisms. The importance of viewing the data of individual organisms was pointed out in the earlier discussion of averaging.

DESIGNS WITH REPEATED MEASURES AND TWO TREATMENT CONDITIONS

Each subject can get both the experimental and the control treatment (see Figure 5.3). The design is otherwise like a design with matched subjects. But care must be taken to deal with effects of order. We will postpone discussing the problem of such effects until after we have dealt with designs having repeated measures on more than one group.

The *t* test for correlated means (see Logistics Box 5.4) is appropriate for data from this type of design, as are the other tests for designs with matched subjects.

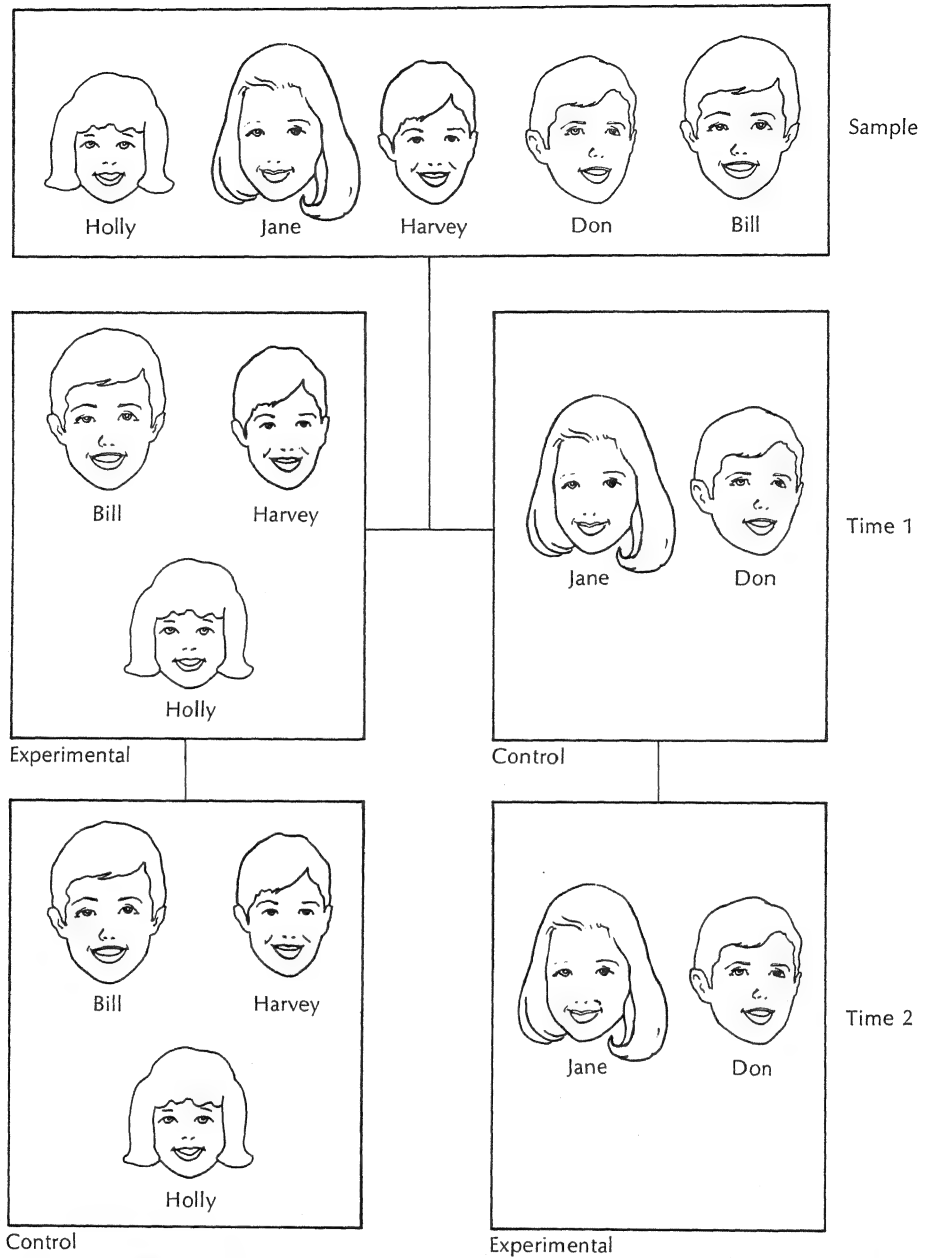



FIGURE 5.3 Design with repeated measures. The same subjects receive all treatments, but in different order.

Logistics Box 5.6: Analysis of variance for randomized blocks design (matched subjects and more than two groups)

1. Find SS_T.
   a. Find C by summing all scores, squaring the result, and dividing by n.

	Treatments A	Treatments B	Treatments C
Block 1	0	1	2
	1	2	3
	0	1	2
	1	2	3
	$\Sigma X_{A1} = 2$	$\Sigma X_{B1} = 6$	$\Sigma X_{C1} = 10$
Block 2	1	2	3
	2	3	4
	1	2	3
	2	3	4
	$\Sigma X_{A2} = 6$	$\Sigma X_{B2} = 10$	$\Sigma X_{C2} = 14$

$$\Sigma X_i = 2 + 6 + 10 + 6 + 10 + 14 = 48$$

$$(\Sigma X_i)^2 = 48^2 = 2304$$

$$C = \frac{(\Sigma X_i)^2}{n} = \frac{2304}{24} = 96$$

   b. Square each score, sum the squares, and subtract C.

	A	B	C
Block 1	0	1	4
	1	4	9
	0	1	4
	1	4	9
Block 2	1	4	9
	4	9	16
	1	4	9
	4	9	16
	$\Sigma X_A^2 = 12$	$\Sigma X_B^2 = 36$	$\Sigma X_C^2 = 76$

$$\Sigma X_i^2 = 12 + 36 + 76 = 124$$

$$SS_T = 124 - 96 = 28$$

2. Find SS_{blocks}.
   a. Add all the scores for block 1 and square the result; repeat for block 2.

      Block 1: $2 + 6 + 10 = 18;\ 18^2 = 324$

      Block 2: $6 + 10 + 14 = 30;\ 30^2 = 900$

   b. Divide each square by the number of scores on which it is based, sum the results, and subtract C.

$$\frac{324}{12} + \frac{900}{12} = 102$$

$$SS_{blocks} = 102 - 96 = 6$$

3. Find $SS_{treatments}$.
   a. Add all the scores for each treatment and square the results.

      A: $0 + 1 + 0 + 1 + 1 + 2 + 1 + 2 = 8;\ 8^2 = 64$

      B: $1 + 2 + 1 + 2 + 2 + 3 + 2 + 3 = 16;\ 16^2 = 256$

      C: $2 + 3 + 2 + 3 + 3 + 4 + 3 + 4 = 24;\ 24^2 = 576$

   b. Divide each square by its corresponding n, sum the results, and then subtract C.

$$SS_{treatments} = \left(\frac{64}{8} + \frac{256}{8} + \frac{576}{8}\right) - 96 = 16$$

4. Find SS for interaction of treatments × blocks.
   a. Square the sums for each treatment group of each block $(\Sigma X_{A1})^2$, $(\Sigma X_{B1})^2$, $(\Sigma X_{C1})^2$, $(\Sigma X_{A2})^2$, $(\Sigma X_{B2})^2$, $(\Sigma X_{C2})^2$, divide each square by the n on which the sum was based, and sum the results.

$$\frac{2^2}{4} + \frac{6^2}{4} + \frac{10^2}{4} + \frac{6^2}{4} + \frac{10^2}{4} + \frac{14^2}{4} = \frac{472}{4} = 118$$

   b. Subtract C, SS_{blocks}, and $SS_{treatments}$ from the result.

$$118 - 96 - 6 - 16 = 0$$

5. Find SS_{error}.

$$SS_{error} = SS_T - SS_{blocks} - SS_{treatments} - SS_{treatments \times blocks}$$
$$= 28 - 6 - 16 - 0$$
$$= 6$$

6. Find the various degrees of freedom.

   For SS_T, df $= n - 1;\ 24 - 1 = 23$

   For SS_{blocks}, df = number of blocks $- 1 = 1$

   For $SS_{treatments}$, df = number of treatments levels $- 1 = 3 - 1 = 2$

   For $SS_{treatments \times blocks}$, df $= df_{treatments} \times df_{blocks} = 2 \times 1 = 2$

   For SS_{error}, df = df for $SS_T - dfSS_{blocks} - dfSS_{treatments} - dfSS_{treatments \times blocks} = 23 - 1 - 2 - 2 = 18$

7. Compute mean squares.

$$MS_{blocks} = \frac{SS_{blocks}}{df_{blocks}} = \frac{6}{1} = 6$$

$$MS_{treatments} = \frac{SS_{treatments}}{df_{treatments}} = \frac{16}{2} = 8$$

$$MS_{treatments \times blocks} = \frac{SS_{treatments \times blocks}}{df_{treatments \times blocks}} = \frac{0}{2} = 0$$

$$MS_{error} = \frac{SS_{error}}{df_{error}} = \frac{6}{18} = 0.33$$

8. Divide the other mean squares by MS_{error} to get F's.

$$F_{blocks} = \frac{MS_{blocks}}{MS_{error}} = \frac{6}{0.33} = 18$$

$$F_{treatments} = \frac{MS_{treatments}}{MS_{error}} = \frac{8}{0.33} = 24$$

$$F_{treatments \times blocks} = \frac{MS_{treatments \times blocks}}{MS_{error}} = \frac{0}{0.33} = 0$$

9. Tabulate the results.

Source	SS	df	MS	F	p
Total	124	23	—	—	—
Blocks	6	1	6	18	< 0.01
Treatments	16	2	8	24	< 0.01
Treatments × blocks	0	2	0	0	N.S.
Error	6	18	0.33	—	—

DESIGNS WITH REPEATED MEASURES AND MORE THAN TWO TREATMENT CONDITIONS

If more than two treatment conditions are used with repeated measures on the same subjects, the design is usually called a *treatments by subjects design.*

The analysis of variance for treatments by subjects designs is explained in Logistics Box 5.7.

S_1: $1 + 3 + 5 = 9,\ 9^2 = 81$ $\qquad 81 + 100 + 100 + 49 = 330$

S_2: $2 + 3 + 5 = 10,\ 10^2 = 100$

S_3: $2 + 4 + 4 = 10,\ 10^2 = 100$ $\qquad \frac{330}{3} = 110$

S_4: $1 + 2 + 4 = 7,\ 7^2 = 49$

b. Subtract C from the result of 3a.

$$SS_{subjects} = 110 - 108 = 2$$

4. Find SS_{error}.

$$SS_{error} = SS_T - SS_{treatments} - SS_{subjects}.$$
$$= 22 - 18 - 2$$
$$SS_{error} = 2$$

5. Determine degrees of freedom.

df for $SS_T = N - 1 = 12 - 1 = 11$

df for $SS_{treatments}$ = number of treatments $- 1 = 3 - 1 = 2$

df for $SS_{subjects}$ = number of subjects $- 1 = 4 - 1 = 3$

df for $_{error} = dfSS_T - df_{treatments} - df_{subjects} = 11 - 2 - 3 = 6$

6. Divide $SS_{treatments}$ by $df_{treatments}$ and SS_{error} by df_{error}.

$$MS_{treatments} = \frac{SS_{treatments}}{df_{treatments}} = \frac{18}{2} = 9 \qquad MS_{error} = \frac{SS_{error}}{df_{error}} = \frac{2}{6} = 0.33$$

7. Find $F_{treatments}$ by dividing $MS_{treatments}$ by MS_{error}.

$$F_{treatments} = \frac{MS_{treatments}}{MS_{error}} = \frac{9}{0.33} = 27.27$$

8. Look up the p related to that F for $df_{treatments}$ (= 2) and df_{error} (= 6).

F for $p = 0.05 = 5.14$ at df = 2 and 6; so reject the null hypothesis.

9. Tabulate the analysis.

Source	SS	df	MS	F	p
Total	22	11	—	—	—
Subjects	2	3	—	—	—
Treatments	18	2	9	27.27	<0.001
Error	2	6	0.33	—	—

METHODS OF DEALING WITH EFFECTS OF ORDER

Order effects must be dealt with when repeated measures are taken on the same subjects. Suppose you wanted to compare the effects on hand steadiness of large doses of alcohol with the effects of equally large doses of milk. You have a device permitting accurate measurement of hand steadiness. You might give the alcohol on Day 1 and then measure hand steadiness; and you might give the milk on Day 2 with the hand-steadiness measure repeated again. If alcohol and milk were administered in that order, the result might be that milk impairs hand steadiness more than alcohol. But the really important disrupter of hand steadiness might well be a hangover from the alcohol. The order of presentation of the two treatments, milk and alcohol, is likely to make a great difference in the experimental outcome. If milk were presented first we would hardly expect it to impair hand steadiness.

Randomization

There are various control techniques for determining the effects of the order of presenting treatments. For example, we can *randomize* order of presentation. In the alcohol versus milk example, randomization of order of presentation would mean that the experimenter would decide for each subject by some random process whether milk or alcohol would be given first. Randomization means that treatments are as likely to occur in one order as in another. It also means that presenting a treatment in one position for a given subject (say, milk first) has no influence on whether the same treatment occurs in any other position (say, milk second) for another subject. This means that the experimenter does not follow a "balancing" rule that milk must be given first a certain proportion of the time. We assume that a truly random process will lead in the long run to a fairly even balance of the various orders of presentation. Randomization thus has the disadvantage that imbalances in order of presentation might occur simply on a chance basis. This is so especially if the number of treatment presentations is small. Randomization will even things out in the *long run,* but only if the experiment is very extensive. It is even possible that a purely random procedure will lead to a situation in which a given experimental treatment always comes first, just as it is possible for someone to draw four aces in a row from a deck of cards without cheating.

Counterbalancing

In order to avoid such imbalances, *counterbalancing* is often used instead of randomization. Counterbalancing means that the experimenter makes sure that the various possible presentation orders occur equally often. With two treatments, an experimental and a control treatment, the design would be counterbalanced if the experimenter

required that the experimental treatment come first on half of the occasions and that the control condition occur first on the other half. By counterbalancing, any effect of either treatment's coming first will be present equally in the experimental and control conditions. By looking at the results when a treatment comes first and comparing it to the results when the same treatment comes second, effects of order can be seen.

Procedures like randomization and counterbalancing permit us to *assess* the effects of order in an experiment. They do not *eliminate* such effects. Suppose you find that order of presentation has an important influence on the outcome of your experiment. It may be worth your while to learn how to manipulate those effects directly through experimental techniques. Ideally, you should learn how to *reverse* the effect of whatever treatments are presented early in the experiment. Reversal of the effect would permit returning the experimental subject to the original baseline before presenting any subsequent treatment. If the subject has been truly returned to baseline, order-of-presentation effects will indeed have been eliminated. However, we often do not have a technique at our disposal for returning a subject to baseline. Some treatment effects are irreversible, or virtually so.

Equivalent tests

Since we cannot always return an experimental subject to the original state, experimenters must rely on approximations to the ideal. For example, an experiment sometimes requires that tests, such as IQ tests, be given twice to the same subject, before and after a treatment. Since a previous single test may influence the score when the subject takes it the second time, there is great danger of an order effect. Psychologists have no practical way of erasing a subject's memory of a previous test. But they can offset its effect by using a device known as an *equivalent form* of the test. A large test is devised and then half of the items are selected randomly for deletion. Those taken out can be expected to be equivalent to those remaining. This is because the deleted items were selected randomly without bias. In this way two equivalent tests are available. A subject can be given one form before the experimental treatment and the other after the treatment. Strictly speaking, the subject is not brought back to baseline. For example, there might be some similar features in the two equivalent forms. But the result is a good approximation to the ideal of returning the subject to baseline.

Running subjects to asymptote

Sometimes subjects will show progressive changes in behavior only up to a certain point. They eventually level off. For example, you generally can only go so far in improving accuracy of basketball shots. Then you stabilize at a certain level of accuracy. Sometimes the progressive

growth in performance can show up as an order effect. But if subjects are first trained until little or no further progress occurs, then the order effect is eliminated.

Providing rest intervals

Many variables have effects that go away by themselves over time. Fatigue is like that. So are the effects of many drugs. By spacing out the various treatments appropriately, we can eliminate the effects. In a repeated-measures design, this can mean the elimination of an order effect. It would prevent confounding in the study of the effects of alcohol versus milk mentioned earlier.

The use of stimulus control

Effects of order can often be minimized by bringing behaviors under stimulus control. A subject may be placed on a *multiple schedule of reinforcement.* Let us say we wish to compare a rat's bar-press behavior under continuous reinforcement (a reinforcement for every bar press) with its behavior under extinction (no response reinforced). The nonreinforced periods might have their effect on behavior during the reinforced periods, and vice versa. But a specific stimulus can be associated with each of the two states of reinforcement. "Light on" might indicate continuous reinforcement, and "light off" might indicate extinction. In time the behaviors typical of continuous reinforcement and extinction will tend to occur only in the presence of their appropriate stimuli. The behavior is then said to be under stimulus control. When there is more than one schedule of reinforcement and more than one stimulus, the procedure is called a multiple schedule of reinforcement. Schedules of reinforcement are discussed in more detail later.

The use of stable baselines

Another method is to develop baseline behaviors that resist order effects. Guttman and Kalish (1956) used this technique in their experiments on stimulus generalization. First organisms were trained to respond in the presence of a given stimulus. Guttman and Kalish wanted to measure response rates in the presence of that stimulus and several other more or less similar stimuli. They wanted to do this in the absence of reinforcement. Since the behavior tends to deteriorate in the absence of reinforcement, stimuli presented late in the series would likely be associated with lower response rates. Guttman and Kalish coped with this order effect by using a schedule of reinforcement that is known to result in behavior that disrupts very little in the absence of reinforcement (a variable interval schedule). Thus, by using a baseline behavior that resisted disruption, order effects were minimized.

The use of probes

Our last example will be that of *probing.* This technique resembles the spot-checking procedures most of us have used from time to time. In this instance, behavior is maintained by some stable technique, and the behavior of interest is sampled only periodically. For example, a sensitive baseline behavior might be brought under stimulus control and the controlling stimulus introduced only periodically. It is essentially a probing technique when sleep researchers awaken individuals periodically during the night and ask them whether they were dreaming. The ongoing silence supports sleep, but does not allow access to verbal reports of dreaming. Waking and asking is a stimulus that probes the state of this sleeping behavior by eliciting verbal reports.

Key Ideas Box 5.2: Methods of dealing with effects of order

Designs with repeated measures on the same subjects require controls for carry-over effects from one treatment to the other (effects of order). A partial list of control methods follows. With *randomization* different orders of presentation are assigned randomly to different subjects. This will balance out effects of order in the long run. With *counterbalancing* the different orders are deliberately arranged to be represented equally often in the experiment by varying the order systematically across subjects. *Experimental reversal of effects* is ideal if within the capabilities of the experimenter. It eliminates carry-over. The use of *equivalent tests* can reduce the carry-over that would result if the same test were repeated. Sometimes *running subjects to asymptote* (the point where their behavior has advanced until there are no further changes) is helpful. For effects that dissipate over time, *rest intervals* between treatments eradicate their carry-over. Effects that might normally carry over can be brought under *stimulus control,* so that they only occur in the presence of a given stimulus and no longer spread to other conditions. Establishment of *stable baselines* may reduce carry-over, since they replace variations in behavior over time with stability. *Probing* means that the subjects' ongoing behavior is disturbed minimally by taking brief behavioral samples under the test conditions. This improves stability and can thereby reduce carry-over.

Multifactor designs

In discussing the various things you can do about uncontrolled variables, I mentioned that a good thing is to *manipulate* them. Students measuring the two-point threshold in their experimental psychology laboratory often worry over the effects of pressure. Some of them choose to manipulate pressure. They still work with the original independent variable (distance between the two points of the stimulus). Thus, they have more than one independent variable in a single experiment. This is called a *multifactor* design. Factorial designs are one kind of multifactor design.

A factorial design can be illustrated as follows:

Loading on Apparatus	Distance between the Two Points (Millimeters)
5	2 4 6 8 10 12 14 16
10	
15	
20	
25	

Notice that with this design, every level of the first factor is paired with every level of the second factor. Five grams is presented with each and every distance between the points. So is 10 grams, and so is every other level of pressure. Furthermore, the 2 mm distance is tried at each and every level of pressure. So is the 4 mm distance, the 6 mm distance, and so on. When all the possible pairings are used in an experiment, it is said to be *completely crossed. A factorial design has more than one independent variable and it is completely crossed.* Factorial designs permit us to study the effects of *limiting conditions.*

Functional relationships inevitably have *limiting conditions.* This means that a given functional relationship will only hold when certain other variables are within specific limits. For example, room temperature must be within certain limits if most behavioral functional relationships are to hold. If the temperature is low or high enough, the organism's behavior will be disrupted, and at certain extremes it will even die.

Normally, we simply take it for granted that there are limiting conditions on our functional relationships. But the fact that there are such limiting conditions implies that variables do not function in isolation but act together to produce an effect. If this is so, there might be some value in seeing how they interact, by manipulating more than one independent variable at a time. By determining whether and to what extent such interactions occur, an experimenter gets a good estimate of robustness or generality of observed functional relationships.

Suppose we are interested in testing the effectiveness of a given experimental method of producing neurosis in animals. We will naturally run subjects with the experimental treatment and control subjects without the treatment or with some placebo treatment. In assessing the reliability of our results, we will replicate, perhaps by using several subjects under each of the conditions of the experiment. Now the question is whether we should replicate under the same conditions for all subjects. That would certainly seem to fit customary scientific procedures.

An experimenter choosing this strategy would keep the treatments as similar as possible across subjects. But suppose we think that the relationship between experimenter and experimental subjects helps to determine whether experimental neurosis will occur? There actually *is* evidence that techniques that normally produce neurotic behavior in animals may fail to produce the neurotic patterns when an experimenter has been very friendly with the animal. It may be that the relationship of experimenter and subject is a limiting condition on the functional relationship.

Why not do some of the replications with a neutral experimenter and some with a friendly one? This would be a factorial design.

In outline form, the experimental design would look like Figure 5.4. It would be called a 2 × 2 factorial design. This indicates that there are two independent variables (call them *factors*) and two levels of each of the independent variables. If we add a condition in which there is a hostile experimenter, it becomes a 2 × 3 factorial design. There are still two factors, but now there are two levels of one factor and three levels of the other. The magnitude of the numbers indicates the number of levels of the factors.

We indicate how many factors by having more numbers. For example, adding a drug versus saline treatment as a third independent variable makes the design a 2 × 2 × 2 factorial design. (This assumes

	Neurosis Treatment	Placebo Conditioning Treatment
Friendly Experimenter	A	B
Neutral Experimenter	C	D

FIGURE 5.4 A 2 × 2 factorial design. See text for explanation.

we are leaving out the "hostile experimenter" level of one factor. Including it would give us a 2 × 2 × 3 factorial design.)

When an interaction occurs, this means that the effect of one variable differs as a function of the level of another variable. A good example of interaction came from a recent experience of mine. When radial steel-belted tires first came to my attention, I went out and bought the best I could find. The promised superior handling properties and wear seemed to make it a worthwhile expenditure. Later, I read in a newspaper article that radial steel-belted tires wear better than regular tires only provided wheels are kept in good alignment; they do not hold up at all well when wheels are out of alignment. Unfortunately, I rarely take care of my car, and the wheels are often out of alignment. Thus, radial steel-belted tires were a poor investment for me.

To state the case in terms of interaction, we would say that there is an interaction between wheel alignment and the performance of radial steel-belted tires. At one level of the alignment factor, the "out-of-alignment" level, the radial steel-belteds do poorly. At another level of the alignment factor, the "in-alignment" level, they do well.

SOME TYPES OF FACTORIAL DESIGNS

We have said that the number of levels of factors can be as large as the experimenter wishes. The number of factors can also be extended indefinitely. Limitations on how large these numbers can be are practical. They have to do with the massiveness of the experiment and the difficulties an experimenter might have in comprehending the meaning of an extremely complex design.

The number of factors and the number of levels within factors are important in identifying a given factorial design. But these designs can also differ in another way. They may have repeated measures within subjects on one or more of the factors, or they may be completely randomized. Some factorial designs are *completely randomized.* Some are *completely repeated.* Others are *mixed,* with independent subjects on some factors and repeated measures on others.

Let's go back to the experiment on inducing neuroses in dogs. A completely randomized factorial design would look like the following:

TABLE 5.4 Completely Randomized Factorial Design

Neurosis Treatment		Placebo Treatment	
Friendly Experimenter	*Neutral Experimenter*	*Friendly Experimenter*	*Neutral Experimenter*
Gordon	Blacky	King	Jennifer
Daisy	Mortimer	Pushkin	Bella
Sissy	Fido	Spot	Vic

All possible pairings of the two factors are included. There are two levels of the "friendliness" factor and two levels of the "neurosis treatment" factor. Thus there are four possible pairings. These are at the head of the table. Look at Gordon. He receives the neurosis treatment and a friendly experimenter. He does not receive any of the other possible pairings of levels. This is true of every other dog in the experiment. Each receives one and only one pair of levels of the two factors.

Contrast this with the following design, which is completely repeated.

TABLE 5.5 Completely Repeated Factorial Design

Neurosis Treatment		Placebo Treatment	
Friendly Experimenter	*Neutral Experimenter*	*Friendly Experimenter*	*Neutral Experimenter*
Gordon	Gordon	Gordon	Gordon
Daisy	Daisy	Daisy	Daisy
Sissy	Sissy	Sissy	Sissy

Each subject receives every possible pairing of levels of the two factors. A mixed factorial design might look like this:

TABLE 5.6 Mixed Factorial Design

Neurosis Treatment		Placebo Treatment	
Friendly Experimenter	*Neutral Experimenter*	*Friendly Experimenter*	*Neutral Experimenter*
Gordon	Gordon	Blacky	Blacky
Daisy	Daisy	Mortimer	Mortimer
Sissy	Sissy	Fido	Fido

Consider Gordon again. This time he receives both levels of the "friendliness" factor. He gets both the friendly experimenter and the neutral experimenter. But he receives only one level of the "neurosis treatment" factor. He is not repeated on that factor.

Incidentally, we would say that with the completely repeated design, subjects are *completely crossed.* If you think of subjects as another factor, every level of it is paired with every level of the other factors.

ANALYSIS OF DATA FROM FACTORIAL DESIGNS

The method of analyzing data from factorial designs depends on such things as the number of factors and whether the design is completely randomized, completely repeated, or mixed. Computational procedures for such analyses are given in Logistics Boxes 5.8, 5.9, and 5.10.

Logistics Box 5.8: Computation of analysis for completely randomized factorial design for two factors and any number of levels

1. Find SS_T.
   a. Find C by summing all the scores, squaring the result, and dividing by the total number of measures.

	Factor A	
	Level 1	Level 2
Factor B Level 1	S_1 0	S_9 5
	S_2 1	S_{10} 4
	S_3 2	S_{11} 5
	S_4 3	S_{12} 4
Level 2	S_5 1	S_{13} 6
	S_6 2	S_{14} 5
	S_7 2	S_{15} 6
	S_8 3	S_{16} 5

$0 + 1 + 2 + 3 + 1 + 2 + 2 + 3 + 5 + 4 + 5 + 4 + 6 + 5 + 6 + 5 = 54$

$$C = \frac{(\Sigma X_i)^2}{N} = \frac{54^2}{16} = 182.3$$

   b. Square each score, sum the squares, then subtract C.

$$0^2 + 1^2 + 2^2 + 3^2 + 1^2 + 2^2 + 2^2 + 3^2 + 5^2 + 4^2 + 5^2 + 4^2 + 6^2 + 5^2 + 6^2 + 5^2 = 236$$

$$SS_T = 236 - 182.3 = 53.7$$

2. Find $SS_{\text{factor A}}$ (SS_A).
   a. Sum the scores within level 1 of factor A across levels of factor B, and square the result and divide by the number of scores on which the sum is based. Repeat for level 2 of factor A; then add the resulting sums.

Level 1: $0 + 1 + 2 + 3 + 1 + 2 + 2 + 3 = 14$, $14^2 = 196$

Level 2: $5 + 4 + 5 + 4 + 6 + 5 + 6 + 5 = 40$, $40^2 = 1600$

$$\frac{196}{8} + \frac{1600}{8} = 224.5$$

   b. Subtract C from the result of 2a.

$$SS_{\text{factor A}} = 224.5 - 182.3 = 42.2$$

3. Find $SS_{\text{factor B}}$ (SS_B).
   a. Sum the scores for level 1 within factor B across levels of factor A, square the result and divide by the number of scores on which the sum is based. Repeat for level 2 of factor B, then add the resulting sums.

Level 1: $0 + 1 + 2 + 3 + 5 + 4 + 5 + 4 = 24$, $24^2 = 576$

Level 2: $1 + 2 + 2 + 3 + 6 + 5 + 6 + 5 = 30$, $30^2 = 900$

$$\frac{576}{8} + \frac{900}{8} = 184.5$$

b. Subtract C from 3a.

$$SS_{\text{factor B}} = 184.5 - 182.3 = 2.2$$

4. Find $SS_{\text{factor A} \times \text{factor B}}$ ($SS_{A \times B}$).
   a. Sum the scores for each of the four subgroups, square each sum and divide each sum by the number of scores on which it is based.

   1,1: $0 + 1 + 2 + 3 = 6$, $6^2 = 36$

   1,2: $5 + 4 + 5 + 4 = 18$, $18^2 = 324$

   2,1: $1 + 2 + 2 + 3 = 8$, $8^2 = 64$

   2,2: $6 + 5 + 6 + 5 = 22$, $22^2 = 484$

$$\frac{36}{4} + \frac{324}{4} + \frac{64}{4} + \frac{484}{4} = 227$$

   b. Subtract C, $SS_{\text{factor A}}$, and $SS_{\text{factor B}}$ from 4a.

$$SS_{A \times B} = 227 - 182.3 - 42.2 - 2.2$$

$$SS_{A \times B} = 0.3$$

5. Find SS_{error}.

$$SS_{\text{error}} = SS_T - SS_{\text{factor A}} - SS_{\text{factor B}} - SS_{A \times B}$$

$$= 53.7 - 42.2 - 2.2 - 0.3$$

$$SS_{\text{error}} = 9.0$$

6. Find degrees of freedom.

   $df_{SS_T} = N - 1 = 16 - 1 = 15$

   $df_A =$ number of levels of factor A $- 1 = 1$

   $df_B =$ number of levels of factor B $- 1 = 1$

   $df_{A \times B} = df_A \times df_B = 1 \times 1 = 1$

   $df_{\text{error}} = df_{SS_T} - df_A - df_B - df_{A \times B} = 15 - 1 - 1 - 1 = 12$

7. Divide each SS by its df to get the mean squares. SS_T is not needed.

$$MS_A = \frac{SS_A}{df_A} = \frac{42.2}{1} = 42.2$$

$$MS_B = \frac{SS_B}{df_B} = \frac{2.2}{1} = 2.2$$

$$MS_{A \times B} = \frac{SS_{A \times B}}{df_{A \times B}} = \frac{0.3}{1} = 0.3$$

$$MS_{error} = \frac{SS_{error}}{df_{error}} = \frac{9.0}{12.0} = 0.75$$

8. Divide MS's by MS_{error} to get the F's, look up p's for each F with df for numerator and denominator.

$$\frac{MS_A}{MS_{error}} = \frac{42.2}{0.75} = 56.26$$

$$\frac{MS_B}{MS_{error}} = \frac{2.2}{0.75} = 2.93$$

$$\frac{MS_{A \times B}}{MS_{error}} = \frac{0.3}{0.75} = 0.4$$

9. Tabulate the analysis.

Source	df	MS	F	p
Total	15	—	—	—
A	1	42.2	56.26	< 0.001
B	1	2.2	2.93	NS
A × B	1	0.3	0.4	NS
Error	12	0.75	—	—

Note: This analysis can be extended to any number of levels of the factors.

A more detailed computational handbook has been provided by Bruning and Kintz (1968). It contains "cookbook" accounts of analyses for various other statistics as well. For a deeper understanding of the methods, I suggest Keppel (1973).

Logistics Box 5.9: Computation of analysis for completely repeated factorial design for two factors and any number of levels

1. Find SS_T.
   a. Find C by summing all the scores, squaring the result, and dividing by the total number of measures.

	A_1, B_1	A_1, B_2	A_2, B_1	A_2, B_2
S_1	0	5	1	6
S_2	1	4	2	5
S_3	2	5	2	6
S_4	3	4	3	5
	$\Sigma X_{A_1B_1} = 6$	$\Sigma X_{A_1B_2} = 18$	$\Sigma X_{A_2B_1} = 8$	$\Sigma X_{A_2B_2} = 22$

$$\Sigma X_i = 6 + 18 + 8 + 22 = 54$$

$$C = \frac{(\Sigma X_i)^2}{N} = \frac{54^2}{16} = 182.3$$

   b. Square each score, sum the squares, then subtract C.

$0^2 + 1^2 + 2^2 + 3^2 + 5^2 + 4^2 + 5^2 + 4^2 + 1^2 + 2^2 + 2^2 + 3^2 + 6^2 + 5^2 + 6^2 + 5^2 = 236$

$SS_T = 236 - 182.3 = 53.7$

2. Find $SS_{subjects}$.
   a. Sum scores for each subject, square each resulting sum, and divide by the number of measures on which it was based. Add up the results.

S_1: $0 + 5 + 1 + 6 = 12,\ 12^2 = 144$
S_2: $1 + 4 + 2 + 5 = 12,\ 12^2 = 144$
S_3: $2 + 5 + 2 + 6 = 15,\ 15^2 = 225$
S_4: $3 + 4 + 3 + 5 = 15,\ 15^2 = 225$

$$\frac{144}{4} + \frac{144}{4} + \frac{225}{4} + \frac{225}{4} = 184.5$$

   b. Subtract C from the result of 2a.

$$SS_{subjects} = 184.5 - 182.3 = 2.2$$

3. Find $SS_{factor\ A}$ (SS_A).
   a. Sum scores within level 1 of factor A, ignoring levels of factor B. Square the sum and divide by the number of measures on which it was based. Repeat for level 2 of factor A. Finally, add up the resulting numbers.

A_1: $0+1+2+3+5+4+5+4=24,\ 24^2=576$
A_2: $1+2+2+3+6+5+6+5=30,\ 30^2=900$

$$\frac{576}{8}+\frac{900}{8}=184.5$$

b. Subtract C from 3a.

$$SS_A = 184.5 - 182.3$$
$$SS_A = 2.2$$

4. Find $SS_{\text{factor B}}$ (SS_B).
   a. Sum within level 1 of factor B, ignoring levels of factor A. Square the sum and divide by the number of measures on which it was based. Repeat for level 2 of factor B. Finally, add up the resulting numbers.

B_1: $0+1+2+3+1+2+2+3=14,\ 14^2=196$
B_2: $5+4+5+4+6+5+6+5=40,\ 40^2=1600$

$$\frac{196}{8}+\frac{1600}{8}=224.5$$

   b. Subtract C from 4a.

$$SS_B = 224.5 - 182.3 = 42.2$$

5. Find $SS_{A\times B}$, the interaction of factor A and factor B.
   a. Square $\Sigma X_{A_1B_1}$, $\Sigma X_{A_1B_2}$, $\Sigma X_{A_2B_1}$, and $\Sigma X_{A_2B_2}$, divide each by the number of measures on which it was based, and sum the results.

$$\frac{6^2}{4}+\frac{18^2}{4}+\frac{8^2}{4}+\frac{22^2}{4}=227$$

   b. Subtract C, SS_A, and SS_B from 5a.

$$SS_{A\times B} = 227 - 182.3 - 2.2 - 42.2$$
$$SS_{A\times B} = 0.3$$

6. Find $SS_{\text{error/A}}$, the sum of squares for error of factor A.
   a. For each subject, sum across levels of B within level 1 of A ($A_1B_1 + A_1B_2$ for S_1, $A_1B_1 + A_1B_2$ for S_2, and so on). Square each resulting sum and divide by the number of measures on which it was based. Repeat within level 2 of A and sum the results.

	$0+5=5$		$1+6=7$
Level A_1:	$1+4=5$	Level A_2:	$2+5=7$
	$2+5=7$		$2+6=8$
	$3+4=7$		$3+5=8$

$$\frac{5^2 + 5^2 + 7^2 + 7^2 + 7^2 + 7^2 + 8^2 + 8^2}{2} = 187$$

b. Subtract C, $SS_{subjects}$, and SS_A.

$$SS_{error/A} = 187 - 182.3 - 2.2 - 2.2 = 0.3$$

7. Find $SS_{error/B}$, the sum of squares for error of factor B.
   a. For each subject, sum across levels of A within level 1 of B ($A_1B_1 + A_2B_1$ for S_1, $A_1B_1 + A_2B_1$ for S_2, and so on.) Square each resulting sum and divide by the number of measures on which it was based. Repeat within level 2 of B.

	$0 + 1 = 1$		$5 + 6 = 11$
Level B_1:	$1 + 2 = 3$	Level B_2:	$4 + 5 = 9$
	$2 + 2 = 4$		$5 + 6 = 11$
	$3 + 3 = 6$		$4 + 5 = 9$

$$\frac{1^2 + 3^2 + 4^2 + 6^2 + 11^2 + 9^2 + 11^2 + 9^2}{2} = 233$$

   b. Subtract C, $SS_{subjects}$, and SS_B.

$$SS_{error/B} = 233 - 182.3 - 2.2 - 42.2 = 6.3$$

8. Find $SS_{error/A \times B}$, the sum of squares for error of the A × B interaction.

$$SS_{error/A \times B} = SS_T - SS_{subjects} - SS_A - SS_B - SS_{A \times B} - SS_{error/A} - SS_{error/B}$$
$$= 53.7 - 2.2 - 2.2 - 42.2 - 0.3 - 0.3 - 6.3$$
$$SS_{error/A \times B} = 0.2$$

9. Calculate df.

df_{SS_T} = number of measures $- 1 = 16 - 1 = 15$

$df_{SS_{subjects}}$ = number of subjects $- 1 = 4 - 1 = 3$

df_{SS_A} = number of levels of A $- 1 = 2 - 1 = 1$

df_{SS_B} = number of levels of B $- 1 = 2 - 1 = 1$

$df_{SS_{error/A}}$ = df for $SS_{subjects} \times df_{SS_A} = 3 \times 1 = 3$

$df_{SS_{error/B}}$ = df for $SS_{subjects} \times df_{SS_B} = 3 \times 1 = 3$

$df_{SS_{A \times B}} = df_{SS_A} \times df_{SS_B} = 1 \times 1 = 1$

$df_{error/A \times B}$ = df for $SS_{subjects} \times df_{SS_A} \times df_{SS_B} = 3 \times 1 \times 1 = 3$

10. Calculate the mean squares.

$$MS_A = \frac{SS_A}{df_A} = \frac{2.2}{1} = 2.2$$

$$MS_B = \frac{SS_B}{df_B} = \frac{42.2}{1} = 42.2$$

$$MS_{A \times B} = \frac{SS_{A \times B}}{df_{SS_{A \times B}}} = \frac{0.3}{1} = 0.3$$

$$MS_{error/A} = \frac{SS_{error/A}}{df_{error/A}} = \frac{0.3}{3} = 0.1$$

$$MS_{error/B} = \frac{SS_{error/B}}{df_{error/B}} = \frac{6.3}{3} = 2.1$$

$$MS_{error/A \times B} = \frac{SS_{error/A \times B}}{df_{error/A \times B}} = \frac{0.2}{3} = 0.07$$

11. Compute F's.

$$F_A = \frac{MS_A}{MS_{error/A}} = \frac{2.2}{0.1} = 22.00$$

$$F_B = \frac{MS_B}{MS_{error/B}} = \frac{42.2}{2.1} = 20.10$$

$$F_{A \times B} = \frac{MS_{A \times B}}{MS_{error/A \times B}} = \frac{0.3}{0.07} = 4.29$$

12. Look up p values for each F based on df for numerator and denominator in step 11.

$p(F_A) < 0.025$	$p(F_B) < 0.025$	$p(F_{A \times B}) = \text{NS}$
df = 1, 3	df = 1, 3	df = 1, 3

13. Tabulate results:

Source	SS	df	MS	F	p
Total	53.7	15	—	—	—
Between Subjects	4.7	7	—	—	—
A	2.2	1	2.2	5.23	NS
Error between	2.5	6	0.42	—	—
Within Subjects	49.0	8	6.13	—	—
B	42.2	1	42.2	39.07	< .001
A × B	0.3	1	0.3	0.28	NS
Error within	6.5	6	1.08	—	—

Note: This analysis may be extended to any number of levels of the factors. The repeated factor is commonly "trials," in which case the unrepeated factor is regarded as "conditions."

Logistics Box 5.10: Computation of analysis for mixed factorial designs for two factors and any number of levels

1. Find SS_T.
   a. Add all the scores, square the result, and divide by the total number of measures.

	Factor with Independent Subjects			
	A_1		A_2	
Repeated Factor B_1	S_1	0	S_5	1
	S_2	1	S_6	2
	S_3	2	S_7	2
	S_4	3	S_8	3
B_2	S_1	5	S_5	6
	S_2	4	S_6	5
	S_3	5	S_7	6
	S_4	4	S_8	5

$$0+1+2+3+5+4+5+4+1+2+2+3+6+5+6+5=54$$

$$C=\frac{(\Sigma X_i)^2}{N}=\frac{54^2}{16}=182.3$$

   b. Square each score, sum the squares, then subtract C.

$$0^2+1^2+2^2+3^2+5^2+4^2+5^2+4^2+1^2+2^2+2^2+3^2+6^2+5^2+6^2+5^2=236$$

$$SS_T=236-182.3=53.7$$

2. Find the sum of squares between subjects ($SS_{between}$).
   a. Sum each subject's scores, square each sum, add them together and divide by the number of levels of the repeated factor (B).

S_1: $0+5=5$, $5^2=25$ S_5: $1+6=7$, $7^2=49$
S_2: $1+4=5$, $5^2=25$ S_6: $2+5=7$, $7^2=49$
S_3: $2+5=7$, $7^2=49$ S_7: $2+6=8$, $8^2=64$
S_4: $3+4=7$, $7^2=49$ S_8: $3+5=8$, $8^2=64$

$$\frac{25+25+49+49+49+49+64+64}{2}=187$$

   b. Subtract C from 2a.

$$SS_{between}=187-182.3=4.7$$

3. Find $SS_{factor\ A}$ (nonrepeated factor).
   a. Sum scores within levels of factor A (across levels of factor B). Square the resulting sums, divide by the number of measures on which they were based, and sum the results.

A_1: $0 + 1 + 2 + 3 + 5 + 4 + 5 + 4 = 24,\ 24^2 = 576$

A_2: $1 + 2 + 2 + 3 + 6 + 5 + 6 + 5 = 30,\ 30^2 = 900$

$$\frac{576 + 900}{8} = 184.5$$

   b. Subtract C from 3a.

$$SS_A = 184.5 - 182.3$$
$$= 2.2$$

4. Find $SS_{error/between}$.

$$SS_{error/between} = SS_{between} - SS_A$$
$$= 4.7 - 2.2$$
$$= 2.5$$

5. Find SS_{within}.

$$SS_{within} = SS_T - SS_{between}$$
$$= 53.7 - 4.7$$
$$= 49.0$$

6. Find SS_B (the repeated factor).
   a. Sum scores within levels of B (across levels of A). Square the resulting sums, divide by the number of measures on which they were based, and sum the results.

B_1: $0 + 1 + 2 + 3 + 1 + 2 + 2 + 3 = 14,\ 14^2 = 196$

B_2: $5 + 4 + 5 + 4 + 6 + 5 + 6 + 5 = 40,\ 40^2 = 1600$

$$\frac{196 + 1600}{8} = 224.5$$

   b. Subtract C from 6a.

$$SS_B = 224.5 - 182.3$$
$$= 42.2$$

7. Find $SS_{A \times B}$.
   a. Sum scores for each subcondition (A_1B_1, A_1B_2, A_2B_1, A_2B_2),

square each result, divide by the number of measures on which it is based, and sum the results.

A_1B_1: $0 + 1 + 2 + 3 = 6$
A_1B_2: $5 + 4 + 5 + 4 = 18$
A_2B_1: $1 + 2 + 2 + 3 = 8$
A_2B_2: $6 + 5 + 6 + 5 = 22$

$$\frac{6^2}{4} + \frac{18^2}{4} + \frac{8^2}{4} + \frac{22^2}{4} = 227$$

b. Subtract C, SS_A, and SS_B from 7a.

$$SS_{(A \times B)} = 227 - 182.3 - 2.2 - 42.2 = 0.3$$

8. Find $SS_{error/within}$.

$$SS_{error/within} = SS_{within} - SS_B - SS_{A \times B} = 49.0 - 42.2 - 0.3$$

$$SS_{error/within} = 6.5$$

9. Compute df's.

df_{SS_T} = total number of measures $- 1 = 16 - 1 = 15$

$df_{SS_{between}}$ = total number of subjects $= 1 = 8 - 1 = 7$

df_{SS_A} = number of levels of A $- 1 = 2 - 1 = 1$

$df_{SS_{error/between}}$ = df for $SS_{between} - df_{SS_A} = 7 - 1 = 6$

$df_{SS_{within}}$ = df for SS_T − df for $SS_{between} = 15 - 7 = 8$

df_{SS_B} = number of levels of B $- 1 = 2 - 1 = 1$

$df_{SS_{A \times B}}$ = df for $SS_B \times df_{SS_A} = 1 \times 1 = 1$

$df_{SS_{error/within}}$ = df for $SS_{within} - df_{SS_B} - df_{SS_{A \times B}} = 8 - 1 - 1 = 6$

10. Compute mean squares.

$$MS_A = \frac{SS_A}{df_A} = \frac{2.2}{1} = 2.2$$

$$MS_{error/between} = \frac{SS_{error/between}}{df_{error/between}} = \frac{2.5}{6} = 0.42$$

$$MS_{within} = \frac{SS_{within}}{df_{within}} = \frac{49.0}{8} = 6.13$$

Key Ideas Box 5.3: Summary of Experimental Designs

Name of Design	Method of Assigning Subjects	Number of Conditions
Completely Randomized (Independent Subjects)	Subjects assigned randomly to conditions. No subject in more than one condition.	2–*n*
Matched Subjects Randomized Blocks	Matched subjects assigned to conditions. Assignment random within the constraint that they be matched.	2–*n*
Repeated Measures Single Subjects; Within Subjects; Treatments by Subjects	Same subjects given each treatment with order of presentation controlled.	2–*n*
Completely Randomized Factorial	More than one independent variable. All possible combinations of levels of the variables represented. Each subject randomly assigned to one combination.	4–*n*
Completely Repeated Factorial	More than one independent variable. All possible combinations of levels of the variables represented. All pairings of levels given to each subject.	4–*n*
Mixed Factorial	More than one independent variable. All possible combinations of levels of the variables represented. Completely randomized on some variables and repeated measures within subjects on others.	4–*n*

UNEQUAL NUMBERS OF SUBJECTS

It is best to have equal numbers of subjects in the various treatment conditions. If inequalities are due to loss of subjects resulting from failure to complete the experiment, the resulting nonrandomness may result in confounding. Furthermore, statistical tests tend to be

Characteristics	Typical Analyses Two Groups	 More than Two Groups
High noise level due to the individual differences' inclusion in error of measurement. Needs larger number of subjects than less noisy designs.	*t* test for independent means; Mann-Whitney *U*	Analysis of variance; Kruskall-Wallis one-way analysis of variance.
Reduced noise due to matching on pre-measure. Pre-measure must be positively correlated with dependent variable.	*t* test for correlated means; Wilcoxon Matched Pairs Signed Rank test; Sign Test.	Analysis of variance; Friedman 2-way analysis of variance.
Reduced noise due to eradication of individual differences in error estimate. Must cope with effects of order of presentation, carry-over from one treatment to the other.	*t* test for correlated means; Wilcoxon Matched Pairs Signed Ranks test; Sign Test.	Analysis of variance; Friedman 2-way analysis of variance.
Factorial designs show variations in functional relationships of one variable when levels of another variable change (interactions), establish generality.	——	Analysis of variance.
Factorial designs show variations in functional relationships of one variable when levels of another variable change (interactions), establish generality.	——	Analysis of variance.
Factorial designs show variations in functional relationships of one variable when levels of another variable change (interactions), establish generality.	——	Analysis of variance.

insensitive to violations of their assumptions *except* when different numbers of subjects are assigned to different conditions. Since variance tends to vary with *n*, experiments with unequal *n*'s tend to produce violations of the assumptions while being sensitive to those violations.

Key Ideas Box 5.4: Unprotected and planned comparisons

If more than one comparison between two groups (such as the *t* test) is done, the alpha level must be made more stringent if the rate of type II errors is to be held constant. Further, an analysis of variance should be done to assure that the findings are, over-all, reliable before individual comparisons are done. An exception is the case of comparisons planned prior to the experiment.

ANTICIPATING THE OPERATION OF LEVINSON'S LAW[2]

Levinson's Law states, "If anything can possibly go wrong, it will." (Levinson, 1967.) The experimenter who fails to operate with an eye to Levinson's Law faces dire consequences. To illustrate, a colleague of mine was directing research on animal learning for almost a year before he learned that two of the three people running the animals were defining an "error" incorrectly. The research director was a highly seasoned experimenter, and well aware of Levinson's Law, but he permitted himself to relax and fall into a fatal optimism.

People *will* fail to understand instructions or *will* perceive what are thought to be agreed-upon procedures incorrectly. Apparatus *will* break down, or worse, malfunction to give spurious data. Human subjects *will* fail to show up, and *will* do unexpected things not anticipated in the procedure. Animal subjects *will* fall ill at critical moments in the experiment or find ways to confound the experimenter's intended procedure.

I once worked for days behaviorally shaping a cat to press a bar for milk reward, only to have it discover a technique for dipping its paw in the milk reservoir and licking ample quantities of milk from it. I have had monkeys escape from a complex automated apparatus and play at tearing the wires up for half an hour before their unauthorized leave from the apparatus was noticed.

Levinson's Law is the dark side of research—an aspect which is seldom revealed in the fine, glossy reports in the journals, and the still

[2]Sometimes called "Murphy's Law." I term it Levinson's law because it was so named in the only printed statement of it I have found (Levinson's *Unafraid Dictionary,* 1967), and in honor of my colleague, Dr. Daniel M. Levinson, whose appreciation of this law is unexcelled.

finer and glossier reports in the textbooks. But the willingness of experimenters to go through this sort of thing indicates how gratifying the rewards are when they come.

A researcher is bound to suffer from the inexorable operation of Levinson's Law from time to time. But you can devise means of coping with it. You should work on projects that you envision as important enough to make the sacrifices worth while. You should keep in mind that many of your research efforts will end in failure. You must simply do more work in anticipation of such failures. You should also *check* and *check* and *double-check* your experiments in all their facets.

A research director who never participates in or observes the project's experiments in operation is asking to be smitten by the mighty hand of Levinson. Be a subject in your own experiments when possible. Even if you cannot use the data, you will learn a great deal. You will probably come away with many new ideas about how the experiment should be run, and will have a good chance at detecting procedural flaws. Whether it is possible to participate as a subject in an experiment or not, it is a good idea for members of research teams to watch each other in operation from time to time. Procedural discrepancies are relatively easy to detect in this way.

The data themselves will often tip off an experimenter that a procedural flaw has intruded. The fiasco mentioned earlier, in which animals were run incorrectly in a learning experiment for almost a year, was detected by looking at the data and noticing that implausibly long runs of errors were produced by many of the subjects. If the experimenters had been sharing their data, presenting them to each other on a regular basis, and summarizing them frequently in lucid ways, the discrepancy would have been detected much earlier.

There are sundry ways to minimize the danger of procedural and logistical error in experiments. Each new experimental situation presents unique possibilities for error and for taking care to avoid error. The most important thing is to leave to chance as few things as possible, never to get overly optimistic about how smoothly things are running, and always to *remember Levinson's Law.*

Key Ideas Box 5.5: Levinson's Law

Levinson's Law states, "If anything can possibly go wrong, it will." (Levinson, 1967). Check and double-check everything. Look over the data for tipoffs that an error has occurred. Be ready for things to go wrong; take them in your stride and cope with them.

WHEN TO PLAN STATISTICAL ANALYSES

It is a good idea to think about statistical analysis while designing the experiment. People often finish their experiment and then begin trying to dig up a statistic to analyze it. This can be very frustrating. Often just looking ahead will cause you to make a small change in the design that makes a very great difference in the ease of doing the statistical analysis.

Study Questions

1. What is experimental design and how does it differ from sampling?
2. What is a completely randomized design?
3. Why randomize?
4. Explain nominal, ordinal, interval, and ratio scales.
5. What statistical tests are used to analyze data from a completely randomized design with two groups? Under what conditions?
6. What statistics are used to analyze data from completely randomized designs with more than two groups? Under what conditions?
7. How serious is it if we violate the assumption of normality of distribution and use the F test?
8. What is a design with matched subjects? What do we gain from matching?
9. Under what conditions is matching wise?
10. What is a yoked control?
11. What statistics are used to analyze data from designs with matched subjects and two groups? Under what conditions?
12. What is a randomized blocks design?
13. How closely correlated must the variables used for blocking and the dependent variable be for blocking to compensate for the lost degrees of freedom?
14. What statistics are used to analyze data from designs with matched subjects and more than two groups? Under what conditions?
15. What is a design with repeated measures on the same subjects?
16. What statistics are used to analyze data from a design with repeated measures on the same subjects and two groups? More than two groups?
17. What is an effect of order?
18. What are the methods used for dealing with effects of order?
19. What is a factorial design?
20. What does it mean to say that a design is completely crossed?
21. What is a factor?

22. What is a level of a factor?
23. Explain the following types of factorial design:
    a. completely randomized
    b. completely repeated
    c. mixed
24. What are the implications of unequal sample sizes in the different treatment conditions of an experiment?
25. Evaluate pilot studies and explain how to sequence experiments.
26. Explain Levinson's Law and tell how to cope with it.

6 LEARNING Fundamentals

Why study learning?

THE POWER OF LEARNING

Larry hurried to keep up with the police officer who led him, handcuffed, into his new home. It was an institution for the mentally retarded. He had been raised in another institution not unlike this one, and he seemed unconcerned about his new situation. "Take the handcuffs off that man," snapped the attendant. "Not until you sign here," retorted the policeman. "He may be dangerous, and I'm responsible for him until you sign."

The place and Larry seemed to suit each other. He walked with the awkward shuffle so common among the brain-damaged; his face sagged and his lower lip was extended as though to prepare for a drool. His interest quickly focused on the glitter of a candy machine in the waiting room. The long past and the indefinite future of confinement seemed not to concern him at all.

It seemed obvious. Larry was the victim of some sort of inherited abnormality, or perhaps of a birth injury that had left him profoundly retarded. He had no family, and there was no place for him except in an institution much like the one he had known from the beginning of his life.

But what seemed obvious was wrong in this case. *Larry was not retarded at all!* At least not in any physical sense. He had *learned* to behave like a retarded person because, for reasons we shall never know, his mother had placed him in an institution for the retarded shortly after birth. All he had ever had an opportunity to learn was abnormal behavior, and so he was virtually indistinguishable from an organically retarded person.

A television special some time ago told the story of how an alert therapist detected Larry's real abilities and taught him to behave in a normal way. Learning had made him retarded and it was through learning that he overcame that retardation.

How we think, perceive, feel, and act depends greatly on our prior learning. And if we want to change, our best hope lies in *learning* new ways of thinking, perceiving, feeling, and acting. So experimental psychologists naturally want to know as much as they can about learning and the laws according to which learning occurs. By understanding the laws of learning, we hope some day to help people to fulfill their enormous potential as human beings.

THE PROMISE OF AN EXPERIMENTAL PSYCHOLOGY OF LEARNING

Frederick K. was a successful lawyer in a small midwestern city. He was in his early thirties, strongly built, and handsome. He had a wife and a three-year-old son. All in all, Fred came across as normal almost to the point of being square. But over the past ten years he had seen perhaps half a dozen therapists.

Fred suffered from a fetish that promised to destroy him socially. He had an uncontrollable urge to smear mucus on handbags. He had been arrested on several occasions for speeding up to female pedestrians on his motorcycle, smearing mucus on their handbags, and attempting to speed away. This was a major source of sexual excitement for him.

Fred heard about a form of behavior therapy, called aversion therapy, that might cure him. It sounded unpleasant, but he was desperate.

"Our method of treatment is based on the principles of respondent and operant conditioning," Dr. Sawyer told Fred. "As you know, many therapists feel that problems like yours can be solved through exploring past experiences you may have had and gaining insight into the basis of your urges. They feel that your symptoms are merely a manifestation of a deeper problem that must be unearthed before you can be cured. We see it as a matter of inappropriate learning.

"Most people react with disgust at the thought of smearing mucus on handbags. You react to it with sexual pleasure. Our task is to replace

the response of pleasure with a response of disgust. This we do through principles of conditioning discovered in laboratories of experimental psychology.

"What we will do is arrange things so that the smearing of mucus is quickly followed by a state of nausea. Pavlov discovered long ago that such an association of two events results in a conditioned response. In other words, you will feel sick at the very thought of smearing mucus on handbags."

"It doesn't sound very pleasant," muttered Fred.

"You're right, Fred, the conditioning procedure will not be pleasant. But only a few trials will be necessary to cure you of the fetish. It's up to you to decide whether it's worth it to you."

Fred had already made up his mind. "I'll do it!" he said. And in a matter of days he was no longer ridden by the fetish.

Many years of study of animal and human learning in very simplified laboratory situations were necessary before methods of treating human disorders were available. Not all of the methods are as unpleasant as aversion therapy. Most of them involve the use of positive reward. Knowledgeable application of reward can permit us to learn with remarkable efficiency.

One method discovered not so long ago in the laboratory is that of *errorless discrimination learning* (Terrace, 1963). Organisms actually learn a discrimination without experiencing any failure at all. This is done by arranging things so that they will spontaneously make the correct response, then making the discrimination more difficult in such gradual steps that the organism always does the right thing.

Suppose a pigeon spontaneously pecks a lighted key. You want it to discriminate red from green keys. Make the light behind the red key very bright. The pigeon will peck it. Then gradually lower the brightness of the red key. This is called *fading* the brightness cue. The pigeon will eventually be cued to peck the red key. It need not experience the failure of reinforcement that goes with pecking the green key.

Similar procedures can be used for training humans to learn discriminations. And most of what we do in school is learning discriminations. To put it in a nutshell, people can benefit immensely from experimental research on learning.

COMMON SENSE IS NOT ENOUGH

Many principles of learning derived from laboratories of experimental research seem like extensions of common sense. Of course people do the things for which they get rewards. Naturally they dislike things associated with punishment. Yet the extent to which these relationships are true cannot be appreciated from common sense alone. For example, few people suspected that responses of the bodily organs

(such as patterns of heartbeats and brain waves) could be controlled through selective reward, but studies of biofeedback show that they often can be.

More striking are findings that run counter to common sense. Take the case of *aversive control*[1] of behavior. We use aversive control very often in our society. If someone does something we dislike, we tend to think of retaliating with punishment in one way or another. We are so accustomed to the use of such unpleasant methods of control that we are hardly aware of them. Try going for a day without answering your friends and acquaintances when they say "Hello." You'll see that they "get mad." What this means is that they start putting aversive consequences on you. Consider also how often you hear people suggest more and more severe punishments when antisocial behaviors come to their attention. We are trained to have a great deal of faith in the power of punishment.

Punishment often works to bring someone's undesirable behavior to a quick stop. The gratifications we gain from such abrupt stoppages probably explain our attachment to punitiveness. However, researchers have identified some interesting situations in which punishment actually helps perpetuate behaviors. *Vicious circle behavior* is a case in point (see Brown, Martin & Morrow, 1964; Martin & Melvin, 1964; Brown, 1969). Suppose a rat is placed in one end of a chamber with a shock grid for a floor, then trained to run to the other end to avoid shock. As long as it runs promptly, it does not get any shock. Of course, it will take some shocks at first, but eventually it will learn to avoid them. Once it has learned this, the behavior may persist for a very long time even if the shock is shut off altogether. You might say that, since it avoids the shocks, it has little opportunity to discover that they are no longer being delivered by the experimenter.

How can we help the rat to get rid of the useless avoidance behavior? Notice that this is not really just a question about rats. Many of our own fear-based behaviors are probably created in much the same way and display the same persistence. Some people seem to live their entire lives running from things that occur rarely or even never.

Why not try punishing the rat for engaging in the previously rewarded running behavior? We could do an experiment comparing rats who simply run until they discover that no more shocks are being given with rats specifically shocked for running (a completely random-

[1]The word "aversive" comes from Latin terms meaning "to turn away from." Control of behavior by environmental events that an organism normally "turns away from" is called "aversive control." It includes *punishment,* in which an aversive event follows a behavior. It also includes *escape,* in which the delivery of a stimulus does not depend on the organism's behavior, but it must respond to terminate the aversive stimulations. A third type of aversive control involves *avoidance* behavior. In that case the response postpones or prevents the aversive event.

ized design). It makes sense to suppose that the punished rats would quit running sooner. But actual experiments of this kind show that they do not! They persist even longer than the unpunished rats.

Research done on a variety of species indicates that punishment, far from being a cure-all for undesirable behavior, often increases the very behavior it is expected to eliminate. Common sense alone will not do. We need research on learning.

Respondent conditioning

An important form of learning is *respondent conditioning.* Respondent conditioning (also known as classical, Pavlovian, or Type S conditioning) was formally discovered[2] as a result of certain incidental observations made by the great Russian psychologist Ivan Petrovitch Pavlov. He was doing research on digestion, for which he won the Nobel Prize. Respondent conditioning is easy to observe. But a scientist of the stature of Pavlov was required to recognize its broader significance.

Pavlov had modified the digestive systems of dogs so that he could make direct observations of the reflexive secretions associated with the digestive process. One operation that he commonly performed is called an "esophagostomy." The esophagus is brought to the surface and attached to the skin of the animal's throat. It is also cut in two, so that everything swallowed by the animal drops out of the upper end of the severed esophagus instead of finding its way to the stomach. It is possible to keep the animal alive and in good health. (Pavlov always insisted that his experimental animals be in good spirits.) Such animals are fed by injecting food into the stomach via the lower end of the severed esophagus. It is easy to observe, in an animal so prepared, that the introduction of food into the mouth will give rise to increased salivation. Furthermore, the animal will salivate at the sight or smell of food, even when it is not actually placed in its mouth. In fact, if the animal caretaker establishes a routine of preparing food in the animal's presence, saliva comes to flow copiously at the earliest signs of preparation—before any food has been introduced. Pavlov, upon observing this sort of phenomenon, realized that the salivary reflex had become conditional on the presentation of previously ineffective nonfood stimuli. The ineffective stimulus had apparently become a signal of the approach of food.

Realizing that he had found a method for making exact, quantitative observations of the development of acquired stimulus-response connections, Pavlov turned all the energies of his laboratory toward the study of these "psychic secretions," or *conditioned reflexes.* Of course

[2]Actually the phenomenon was understood even in the Middle Ages, but Pavlov, who really made it famous, is credited with the discovery.

Ivan Petrovitch Pavlov

the technique for studying conditioned reflexes was considerably more refined than the method that gave rise to the initial, rather casual, observations. In order to get an exact measurement of the dependent variable, salivary flow, a fistula was made in the dog's cheek. This permitted the flow from the salivary glands to be channeled into a tube. The animal, nearly always a dog, was adapted to the experimental situation and trained to remain on a stand to which it was attached by a loose harness (see Figure 6.1). The sound of a buzzer, the beat of a metronome, the sound of bubbling water, or other signals were used as neutral stimuli. A pneumatic device was developed to permit direct introduction of food powder into the mouth. The laboratory constructed for Pavlov permitted excellent control of extraneous stimulation, minimizing "noise levels" in his experiments.

A typical experimental outcome is illustrated in Table 6.1. The experiment, which involved repeated measures on the same subject, was done by Anrep (1920). A tone (S_n) was paired with the presentation of food powder (termed the *unconditioned stimulus,* or *UCS*). Initially, the neutral stimulus (S_n) did not elicit the response of salivation. The UCS (food powder) did elicit a salivary response, of course. This response is termed the *unconditioned response,* or *UCR.* After a number of pairings of S_n with the UCS, salivation was elicited by the previously neutral tone. At this point we speak of the tone as a *conditioned stimulus (CS).* The salivary response that it elicits is a

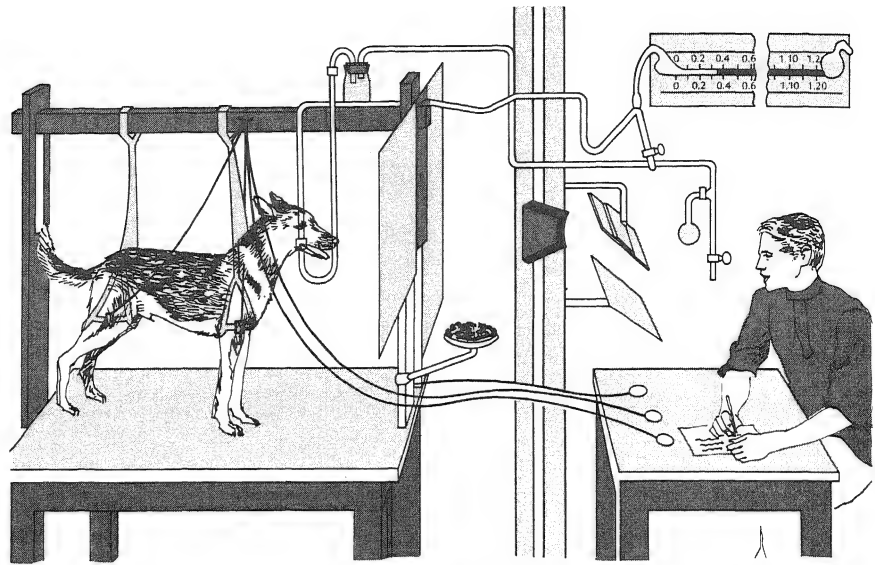



FIGURE 6.1 Pavlov's apparatus for salivary conditioning. The dog is held by a loose harness; a suction cup is connected to an opening made surgically from the salivary gland. Saliva is measured by the device shown above the experimenter. A food dish is on a swiveled shaft, near the center of the figure. The experimenter is in a separate room; he can control stimulus presentation and observe the dog. (From Stagner & Solley, 1970. After Pavlov, 1928.)

conditioned response (CR), and we say that we have established a *conditioned reflex.*

Table 6.1 illustrates several features of the development or *acquisition* of the CR. Note that two quite different observations reflect the acquisition process. The size or quantity of the response was measured (response magnitude). The amount of time elapsing between presentation of the CS and the onset of salivation, called the response *latency,* was also recorded. The latency measure is meaningless if no salivary response occurs, as was the case after only one paired presentation.

TABLE 6.1 A Typical Outcome Obtained in a Respondent-Conditioning Experiment

Number of Paired Stimulations	Response Magnitude (drops of saliva)	Response Latency (seconds)
1	0	
10	6	18
20	20	9
30	60	2
40	62	1
50	59	2

From Anrep, 1920.

TABLE 6.2 Terminology for Stimuli and Responses in a Respondent-Conditioning Experiment

Neutral stimulus (S_n)	An environmental event (such as a tone) that does not normally produce the response the experimenter wishes to condition (such as salivation).
Unconditioned stimulus (UCS)	An environmental event (such as food) that normally produces the response the experimenter wishes to condition (such as salivation).
Unconditioned response (UCR)	The response normally made to the unconditioned stimulus (for example, *salivation* to food, *fear responses* to a clang).
Conditioned stimulus (CS)	A formerly neutral stimulus, which, as a result of conditioning procedures, now produces all or part of a UCR (for example, *tone* produces salivation, *white rat* produces fear).
Conditioned response (CR)	The aspect of a UCR that comes to be produced by a CS as a result of conditioning procedures (for example, *salivation* produced by a tone).

Generally, latency decreased in this experiment as response magnitude increased. A final point is that performance improved up to about 30 trials, after which both types of measure indicate that it leveled off.

THE GENERALITY OF RESPONDENT CONDITIONING

Pavlov's discovery of respondent conditioning would have been of little significance if it were not for the great generality of this kind of conditioning. Chapter 2 pointed out the importance of generality. With few exceptions, virtually any response can be controlled by respondent conditioning. It is highly general. Pavlov used to give a striking demonstration of its power by accustoming a dog to having a particular assistant inject it with apomorphine, which causes sweating, shivering, salivating, vomiting, and eventually an exhausted sleep. While Pavlov gave lectures on respondent conditioning, the dog was present on stage. At a given point, the assistant walked on stage, and the dog produced all the symptoms of apomorphine just at the sight of him. Respondent conditioning can obviously be a whopper variable.

Imagine how our emotions, attitudes, and physiological responses must be influenced by accidental pairings of neutral stimuli with unconditioned stimuli. Such things as phobias, hallucinations, and asthma attacks have been conditioned (Watson & Rayner, 1920; Ellison, 1941; Justesen et al., 1970).

People have learned to control the electrical resistance of their skin and the temperature of their hands in this way. Since a conditioned stimulus which is under voluntary control can be selected, people can gain control of "involuntary" responses by manipulating the condi-

tioned stimuli. An extremely wide array of species throughout the animal kingdom are susceptible to respondent conditioning.

We have looked mainly at examples of types of unconditioned responses that can become conditioned. Of equal interest is the array of neutral stimuli that can become conditioned stimuli. Pavlov showed that electric shock could act as a "neutral" stimulus, even when the current was supplemented by burning and wounding of the skin. This led directly to the realization that pain can be controlled by conditioning, and resulted, among other things, in the Lamaze "psychoprophylactic" method of natural childbirth. One's own inner responses can also be used as neutral stimuli. Thus a person can condition, say, a drop in temperature of the hand with a given word as the neutral stimulus, then control the temperature merely by thinking the word.

Special relationships between CS and UCS

Sometimes a special relationship holds between certain neutral stimuli and a given type of unconditioned stimulus. In an older study, Valentine (1930) showed that it was easy to condition fear responses to

Key Ideas Box 6.1: Respondent conditioning and its generality

Respondent conditioning (also called Pavlovian or classical conditioning) occurs when a neutral stimulus (such as a buzzer) is paired with another stimulus (unconditioned stimulus, such as food) that elicits the response (unconditioned response, such as salivation) to be conditioned. The neutral stimulus comes to elicit the response. The neutral stimulus is then called a conditioned stimulus and the response it elicits is called a conditioned response.

Pavlovian conditioning has great generality. Species ranging from brainless flatworms to human beings are subject to it. Stimuli ranging from painful pinpricks to mere lapses of time can act as conditioned stimuli. Even one's inner thoughts can be so used. Unconditioned stimuli range from simple food or shock to stimuli that produce profound sickness or emotionality.

Traditional experiments indicated that the best interval between neutral stimulus and unconditioned stimulus was 0.5 seconds. This interval is not general across all types of stimuli and responses. We now know that this interval depends on the type of stimuli and responses involved. Very long intervals are effective for gut-related responses and stimuli that relate to taste or smell.

certain things and very hard to condition them to others. Garcia and Ervin (1968) showed that when animals suffer a general sickness (UCS), they display avoidance responses to chemical stimuli (taste and smell), but not to telereceptive stimuli (auditory and visual stimuli). When they suffer body-surface pain, the converse is true. A further interesting finding was that the CS-UCS interval could be very long (even hours) when tastes were linked with a noxious agent. The latter finding is of great importance to behavior therapists, who are likely to use a half-second CS-UCS interval (formerly believed to be an optimal interval) in attempting to condition aversive responses to such things as alcohol consumption (see, for example, Beech, 1969, p. 21).

EXTINCTION

An experimenter can disrupt the established bond between a conditioned stimulus and its response by repeatedly eliciting the response without presenting the unconditioned stimulus. The probability of the response declines rather quickly and steadily with such unreinforced presentations of the CS. For experimental purposes, the question arises "When extinction may be considered complete?" Strictly speaking, it may be almost impossible to tell when there is no longer any bond between CS and CR. For practical purposes, experimenters select an operational definition, a *criterion of extinction,* which is arbitrary. Thus, you might define extinction as "ten consecutive presentations of the CS without the occurrence of the CR."

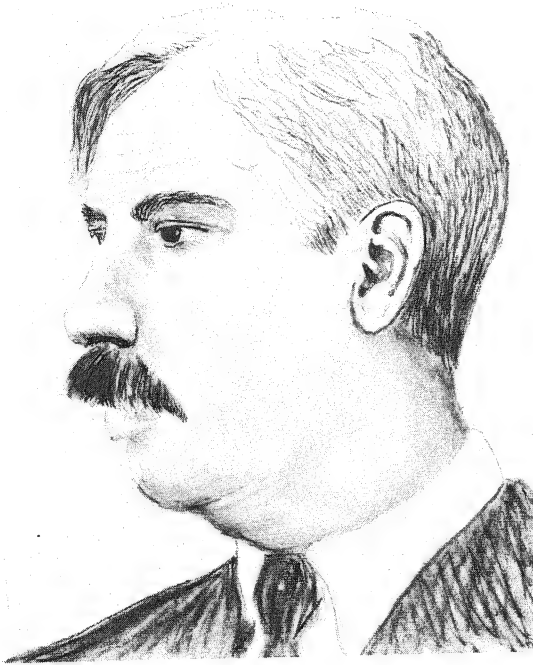
In most cases the operationally accepted notion of extinction does not correspond with the complete removal of all tendency for the organism to make the CR. Various experimental observations reveal this. For example, an extinguished CR will commonly recur if some novel stimulus such as banging a door is introduced into the environment. Pavlov thought that extinction was due to the self-inhibition by the nervous system of its tendency to respond, and he therefore interpreted such recurrences of the extinguished CR with novel stimulation as being due to release from the inhibition. The phenomenon is still called *disinhibition.* A second, related phenomenon is *spontaneous recovery.* Following extinction, an organism will usually show some recovery of the CR after a period of time (say, overnight). Spontaneous recovery is not complete, and the amount of it becomes less and less with repeated re-extinctions.

Operant conditioning

At about the time when Pavlov was conducting his studies of respondent conditioning, Edward Lee Thorndike (1898) was carrying out a series of experiments on animal intelligence, which yielded a phenom-

B. F. Skinner

enon that later (Skinner, 1938) was to be recognized as a second, distinct kind of conditioning—*operant conditioning.* A typical experiment of Thorndike's involved placing a cat in a small box from which it could escape only by pulling a string that dangled from the top of the box. Initially, the cat would make futile attempts to escape—scratching at the walls, meowing, pacing, and so forth—until by chance it happened to hit upon the correct response. Once it pulled the string, the door would open and it would gain its freedom and perhaps a reward of some sort. Later, Thorndike would put the cat back in the box and the escape process would be repeated again, the correct response occurring perhaps a bit more promptly and the irrelevant responses being somewhat reduced. By a gradual process, the cat learned to pull the escape string as soon as it was put in the box. From Thorndike's point of view, the incorrect responses had been stamped out and the correct one stamped in because the incorrect responses were never followed by the satisfaction of escape, whereas the correct responses permitted escape.

There are many points of similarity between the puzzle-box experiment and the Pavlovian experiment. In both cases the animals are

Edward L. Thorndike

faced with stimuli, make responses, and receive some kind of *reinforcement.* But there are important differences. The most important differences are suggested by one of the names for learning of the Thorndikian type—*effect learning.* In Thorndike's experiments the animal learns by trial and error, and by trial and success. The *effect* of responses determines which responses will occur; or, put another way, the occurrence of the reinforcement *depends upon* the occurrence of the correct response. We usually say that reinforcement is *response-contingent.* But in the Pavlovian situation, reinforcement is not response-contingent; the food is delivered whether the animal does or does not salivate. In fact, it never salivates to S_n in the early trials; otherwise the stimulus would not be truly neutral. Operant conditioning is called "operant" because the response *operates* on the environment; it is called "effect learning" because the response that becomes stamped in is the one with the appropriate *effect.* It is also frequently called "instrumental conditioning" because the response is *instrumental* in attaining the reinforcement.

In operant conditioning, effect determines the occurrence of the correct response; in respondent conditioning, the occurrence of the

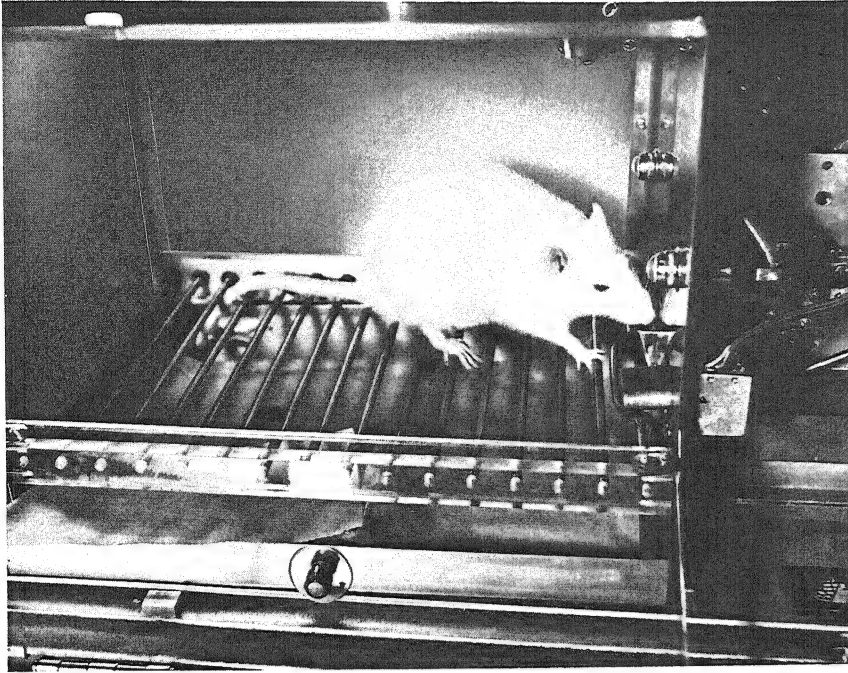



FIGURE 6.2 A free-operant test chamber (Skinner box). The rat is pressing a bar with its left paw while drinking liquid reinforcement from the reinforcement magazine. Another bar is visible to the rat's right. Stimulus lights, which can be used to bring the rat's behavior under stimulus control, are visible on the front wall of the chamber. The rat is standing on grids that can be used to deliver shock.

Why do we call the Skinner box a "free operant" situation? It is "free" because an organism can respond as rapidly as it "likes" without waiting to be put back in a starting position. *Response frequency* or *rate* becomes the convenient measure because of this.

ESTABLISHING AN OPERANT RESPONSE

Usually (but not always) experimenters who use the free-operant technique are not particularly interested in the initial acquisition of the operant response. They prefer to study acquisition from a known baseline rather than from an unknown level (Sidman, 1960). A naive animal placed in a Skinner box may have a variety of responses attached to various aspects of the situation as a result of its prior experience in similar environments, such as the home cage. Some of these responses may aid acquisition of the operant behavior, whereas others (for example, scratching or sleeping) may interfere with it and therefore must be eliminated before the operant response will occur.

Consequently, acquisition measured from an unknown baseline is likely to be highly variable, particularly from one subject to another. It therefore seems preferable to work from a known baseline by, for example, developing the simple bar press first and then measuring the acquisition of some modification of the simple response.

The basic operant response is usually established by the quickest, easiest means possible. A typical procedure is as follows: First, the animal (say, a rat) is placed on cyclical deprivation for a week or two. One might allow him access to food for only 1 hr of every 24 hr. Next, it is placed in the operant chamber and perhaps given some time to explore the new surroundings. Before beginning the actual training procedure, the experimenter may or may not measure the rat's *operant level.* The operant level is the rate of performing the desired response before any specific conditioning procedures have been introduced; that rate can later be compared to the rate attained after various conditioning procedures have been employed. The first step in conditioning the animal is called *magazine training.* The rat learns at this stage that food is available in the chamber, where it is to be found, and that some stimulus such as a click indicates when it is to be found. After it has learned to run promptly from various parts of the chamber to the magazine when it hears the sound of food delivery, *shaping* begins. Shaping means that the subject is reinforced for *successive approximations* to the correct response. We might reinforce it for staying in the half of the chamber nearest the bar, then for staying directly in front of the bar, then for standing on its hind legs in front of the bar, then for touching the bar, and finally for actually pressing the bar. After this, the response is simply *maintained* by continuing to reinforce the response.

The use of shaping procedures in developing an operant behavior is a valuable asset because we are able in this way to teach responses that might otherwise never be developed. It is necessary that the operant level be above zero (that the subject respond at least *sometimes*) before we can deliver a reinforcement. We get around this by initially reinforcing responses that only approximate the correct one but that have a relatively high operant level. In this way we are able to take advantage of the phenomenon of *response variability,* or *response equivalence,* in conditioning an operant behavior. An operant response is defined in terms of the result it accomplishes and not in terms of the muscular movements of which it is composed. For example, an animal may press a bar in many different ways—with its right paw, with its left paw, with its snout—but all these movements are treated as equivalent, as long as they result in depression of the bar. In shaping, the experimenter takes advantage of this equivalence by selectively reinforcing the movements that bring the rat closer to the operant effect ultimately desired. For example, the response of "staying in the front

half of the box" can be accomplished by staying at the right of the bar, or at the left of the bar. If the experimenter selectively reinforces staying in front of the bar, the probability of that type of response component increases, hence making it more likely that "standing up in front of the bar" (a component included under "staying in front of the bar") will occur.

Note the many preparatory details experimenters must attend to before actually getting into their designed experiments. Most of them are taken for granted in published work.

THE GENERALITY OF OPERANT CONDITIONING

Phylogenetic generality

Operant conditioning seems, at least superficially, somewhat more complex than respondent conditioning, and it also appears to fall more readily into a class of behavior that we might loosely call "voluntary." In this regard operant behaviors are to be contrasted with the relatively reflexive respondent behaviors. It would not be unreasonable to infer from this that operant-conditioning principles apply only to those organisms that are relatively high on the phylogenetic scale. However, this inference is incorrect. The available data indicate that respondent and operant conditioning are about equally general in this respect. To list all species that have been successfully conditioned with operant procedures would be pointless. Instead, we will describe a few interesting experiments to illustrate the point with organisms from various phyla. Best (1963) has shown that planaria (flatworms) are capable of learning to go to the correct side of a Y maze when they are rewarded for the correct response by the flooding of the initially dry maze. Actually, Best reports, planaria acquire remarkably complex discriminations and have a tendency to "get bored" in a monotonous environment.

Probably the most widely used species in operant-conditioning studies today is the pigeon. Examples could be selected almost at random, and many of them will be found in various parts of this book. For the present, we will go on to other indications of the generality of operant conditioning.

Operant control of verbal responses in adult humans has been demonstrated repeatedly. The original, and by now classic, experiment in this area is that of Greenspoon (1955). Greenspoon was interested in determining whether the nondirective therapist might inadvertently reinforce certain patterns of verbalization merely by saying "mm-hmm" at appropriate times. He tested this hypothesis by presenting just such stimuli, contingent upon the subject's use of plural nouns, and found that the rate of emission could indeed be controlled in this way.

Generality across types of response

The procedures of operant conditioning have been applied to shape a very wide range of responses. These include such interesting responses as verbal behavior in previously mute psychotics, stuttering, drinking of alcohol on the part of alcoholics, the control of activity of bodily

Logistics Box 6.1: The "galvanic skin response"

The "Galvanic Skin Response" (GSR) has been used a great deal by psychologists as a measure of psychological states such as emotionality and attention, and even as a method of detecting lying.

The GSR is based on the fact that changes in the electrical properties of the skin occur when a person is aroused or stressed. The exact basis of these changes is not fully understood. Activity of sweat glands, muscles, and blood vessels may all make a contribution.

A variety of electrical changes occur in the skin during arousal or stress. There are active changes in *electrical potential* or voltage between two different spots on the skin. Also, the extent to which the skin hinders the flow of an electrical current imposed from outside will vary. This is called the *electrical resistance* of the skin. It is measured in *ohms.* The resistance is greater when an organism is *less* aroused. So it is common to measure the GSR in terms of *conductance,* the reciprocal of resistance ($\text{Conductance} = \frac{1}{\text{Resistance}}$). This goes up during arousal. The unit of measure of conductance is the *mho* (ohm spelled backward).

It is important to distinguish between *basal skin resistance,* which can indicate a person's baseline level of arousal, and the short-range changes of resistance that occur between 0.5 sec and 5 sec after an arousing stimulus is presented. There is confusion in the published literature on just how to apply the term "Galvanic Skin Response." Should we use it for active skin potential changes, for baseline levels of skin resistance, or for the short-range changes in resistance? A new nomenclature has recently been gaining acceptance. The following table shows the new terminology:

Skin potential—the baseline voltage between two skin areas.
Skin potential response—the short-range change of voltage between two skin areas that occurs when a stimulus is presented.
Skin resistance—the baseline electrical resistance of the skin.
Skin resistance response—the change in skin resistance following a stimulus.
Skin conductance—the reciprocal of skin resistance.
Skin conductance response—the reciprocal of skin resistance response.

systems such as the heart, kidney, and peripheral vasculature, and many others.

The basic operant procedure results in the control of virtually any type of response. This is of great practical importance, and has given rise to the therapeutic technique known as *behavioral modification.* Technically, however, the stringent control procedures needed to tell whether control was due to operant conditioning and nothing else have seldom been run. It is not easy to be sure that respondent conditioning, in particular, has not played a role. Great care must be taken if respondent-operant overlap is to be eliminated.

A few years ago great excitement was engendered over research indicating that a wide range of visceral and glandular responses, including such things as heart rate, blood pressure, and urine formation, could be changed through operant conditioning procedures (Miller, 1969). Great care was taken to assure that this was direct operant conditioning of these systems, which had been thought of as solely under "involuntary" control and subject exclusively to respondent conditioning. The rat subjects were even temporarily paralyzed with a drug in order to be sure they were not using voluntary muscles to mediate the internal responses. Paralysis of skeletal muscles is an essential control if we are to conclude that conditioning occurred via the involuntary systems. Otherwise the rats might, for example, speed the heart by struggling or slow it by the profound relaxation of "playing possum." The demonstration seemed clear. The responses came under operant control.

Unfortunately, these results are now in doubt. Miller and his associates cannot replicate the phenomenon reliably (Miller, 1974). The reasons for this are currently a matter for speculation. The search for limiting conditions is still going on. For our purposes, the main point is that it is very, very difficult to be sure we are dealing with "pure" operant conditioning. For most practical purposes such purity is unnecessary. We can permit deliberate confounding. The booming areas of behavioral modification and biofeedback show that a great deal can be accomplished even when deliberate confounding of respondent and operant procedures is tolerated. Nevertheless, we should realize that full understanding of the underlying mechanisms will likely require more refined control procedures.

PROCEDURAL VERSUS FUNCTIONAL DISTINCTION BETWEEN RESPONDENT AND OPERANT CONDITIONING

The distinction between operant and respondent conditioning has been drawn in two very different ways. The two types of conditioning differ with respect to the *procedures* introduced by the experimenter.

With operant conditioning, reinforcement depends on the subject's response. With respondent conditioning, the UCS is delivered according to a program that is independent of the organism's behavior. Thus, the main *procedural distinction* between the two is that reinforcement is response-contingent only in the operant case.

Aside from the procedural differences, some psychologists have argued that operant and respondent conditioning are also functionally different. This means that the two types of conditioning do not follow the same laws. For one thing, it has long been maintained that respondent conditioning alone could control involuntary responses such as those of the gut and the heart.

Currently, many people doubt that operant and respondent conditioning are functionally different. Research on the control of visceral and glandular responses is inconclusive, but has led many to scrutinize the claims of functional differentiation of respondent and operant conditioning. Early reports have at least created a great deal of room for doubt that there is a functional difference. It seems more parsimonious to assume one set of laws until there is solid evidence to the contrary.

Key Ideas Box 6.2: Operant conditioning and its generality

If reinforcement depends on the emission of a response, the resulting conditioning is considered to be of a different type than respondent conditioning. It is called *operant conditioning.* Note that this differs from respondent conditioning in that the neutral stimulus and unconditioned stimulus are, in the respondent case, paired without regard to the organism's behavior.

Operant conditioning also occurs in organisms from flatworms to humans. Behaviors from bar presses to verbal responses, compulsions (such as drinking in alcoholics), and perhaps even biological functions of the organs may be brought under operant control.

It is not at all clear that operant and respondent conditioning work in fundamentally different ways. Even the procedures that identify them as two distinct types of conditioning often overlap in any given situation. For example, a Pavlovian eyeblink response incidentally reduces the noxious stimulus to the eye, creating an unintended operant consequence. And operant reinforcement must be preceded by stimuli in the environment that form respondent conditioned stimuli. This confounding of the two procedures is called *respondent–operant overlap.*

Key Ideas Box 6.3: Eliminating conditioned responses

A conditioned response, once established, may be eliminated in a variety of ways. *Extinction* means that reinforcement is withheld and responses either elicited (respondent conditioning) or allowed to be emitted (operant conditioning) until some arbitrary criterion on nonresponding is met.

Punishment has undesirable side effects, and sometimes increases instead of decreases the target behavior. For example, punishment added to extinction of a response based on avoidance of electric shock can cause a persistence of the response called *vicious circle behavior.*

The phenomenon of *operant–respondent overlap* also suggests commonalities between the two types of conditioning. The procedure intended by the experimenter often turns out not to be the procedure experienced by the subject. When a rat goes down a runway and gets a food reward at the end, this is a case of operant conditioning. It must respond in order to get the food. But incidental to the main procedure, there is a respondent conditioning process going on. Just before reaching the food, the rat is in the presence of stimuli such as the various aspects of the apparatus just outside the goal box. From its point of view there is a pairing of these stimuli with the food, and this is a respondent situation. So both operant and respondent conditioning take place.

Similarly, there may be operant conditioning inherent in certain respondent conditioning procedures. Take the case of eyeblink conditioning. This is one of the most widely used respondent conditioning procedures. A neutral stimulus is paired with a puff of air to the eye. Under such conditions, the eyelid (or, in some species, the "nictitating membrane") shuts in response to the UCS. But when it shuts, it reduces the intensity of the puff of air. Hence, an operant reinforcement is incidentally introduced into the situation. The two procedures are confounded.

A lot of research is currently being done on the question whether operant and respondent conditioning are functionally as well as procedurally different. The research is designed to unconfound the two procedures. For now, we cannot be sure of the outcomes. The principle of parsimony indicates that we should assume only one set of laws, but the data are certainly not all in.

OPERANT REINFORCEMENT

Despite much effort, no uniquely defining property of reinforcers has been identified. A common device is to rely on the *empirical law of effect,* which simply identifies reinforcers as stimuli that, when made contingent on behavior, increase its probability. But there are even difficulties with that view of reinforcement, though it works in the typical case.

Reinforcers may be *primary* (such as food), which means they work without having been made effective through training procedures. They may also be *secondary* or *conditioned.* This means they are made effective through association with primary reinforcement.

Reinforcers may also be *positive* or *negative.* The *introduction* of a positive reinforcer after a response increases the probability of that response. The *withdrawal* of a negative reinforcer after a response results in an increased probability of that response. You can see from the definition that a negative reinforcement is not the same as a punishment. Reinforcers always *increase* the likelihood of responses.

Schedules of reinforcement

The schedule according to which reinforcement is delivered has a great influence on operant responding. For example, the rate of responding is much greater if reinforcement is given for every fiftieth response than if it is given for every response. Schedules of reinforcement are especially important because they produce behavioral baselines that can be used in many different experimental situations. Major schedules of reinforcement and the behaviors that are associated with them are summarized in Table 6.3. A nice, simple source for more detail on schedules of reinforcement is Thompson and Grabowski (1972). The standard reference work is Ferster and Skinner (1957).

Stimulus generalization

Superficially it might seem that the development of a stimulus-response connection by itself accounts for the modification and maintenance of differential response probabilities. But this is actually not enough. Probably no stimulus ever occurs in exactly the same way on two different occasions. For example, the exact angle at which an organism views the stimulus may be slightly different from one occasion to the next, or some aspect of the over-all stimulus complex, either outside or inside the organism, may change. Even if all these things could somehow he held constant, stimuli would change over time. That we can reliably differentiate stimuli occurring at different times (for example, calling them "the first stimulus" and "the second stimulus") is an indication that they are not identical. As the ancient

Greek, Heraclitus, put it, "all things flow," and "you never step into the same stream twice."

If we are to account for the learning phenomenon, it is necessary that we postulate something more than the establishment of "receptor-effector connections" directly involved in the reinforcement process.

We must expand that notion to include not only the exact stimulus that was present during reinforcement, but also other similar stimuli: "The fact is that every reinforcement mediates connections between a very great number of receptor and effector processes in addition to those involved in the reinforcement process and represented in the conventional symbolism $_{S}H_{R}$."[3] (Hull, 1943, p. 183).

Stimuli are often either equivalent or nearly equivalent to each other in their effect on behavior. Put another way, organisms *generalize* from one stimulus to other similar stimuli. This is the *principle of stimulus generalization.* Responses occur in the presence of stimuli that differ considerably from the original stimulus, but that nevertheless have something in common with it. It would be reasonable to ask at this point whether there might not be some systematic relationship between the degree of similarity to the original stimulus and the strength of the response that is evoked. A great deal of attention has been given to this question, and we speak of it as the search for "a gradient of stimulus generalization."

The search for a quantitative law of stimulus generalization has often taken the form of a search for something which Hull referred to as "the true generalization gradient" (Hull, 1943, p. 186). Thus, many investigators have hoped to find a single, precisely quantitative gradient that would apply universally (or at least nearly so). The extent to which this aspiration has been fulfilled may be seen by examining the generality of stimulus generalization. Such functions have not proved to be general.

DESIGNING AN EXPERIMENT TO MEASURE STIMULUS GENERALIZATION

The basic questions experimenters want to answer when they measure stimulus generalization are: "To what extent will the response occur to stimuli not used during training?" "How will the intensity or rate of response vary as stimuli less and less like the original stimulus are used?" They would really like to know the response strengths attached to various stimuli right at the end of training with the original stimulus. Unfortunately, all the methods of measuring stimulus generalization are obtrusive. The process of testing response strengths has a

[3]In the Hullian system, the symbol $_{S}H_{R}$ indicates "habit strength," the strength of the bond between stimulus and response.

TABLE 6.3 Some Major Schedules of Reinforcement

Name of Schedule	Abbreviation	Has or Does Not Have S^d
Continuous Reinforcement	CRF	Either way
Extinction	EXT	Either way
Fixed Ratio	FR	Either way
Variable Ratio	VR	Either way
Fixed Interval	FI	Either way
Variable Interval	VI	Either way
Differential Reinforcement of Low Rates of Responding	DRL	Either way
Differential Reinforcement of High Rates of Responding	DRH	Either way
Chained	Chain	Has S^ds
Tandem	Tand.	No
Multiple	Mult.	Has S^ds
Mixed	Mix	No
Concurrent	Conc.	Either way

Response Contingency	Patterns of Responding
Reinforcement after each response	Moderate rates, not very stable, low resistance to extinction
No responses reinforced	Depends on preceding schedule at first, then rate drops to criterion of extinction
Reinforcement for a fixed number of responses greater than one	Very high rates; pauses occur after reinforcement if ratio is high; resistance to extinction high and in runs of responses about equal to the ratio
Reinforcement after varying numbers of responses with a specified mean number	High stable rates, very great resistance to extinction
Reinforcement for each response provided a fixed interval has passed since the last reinforcement	"Scalloping," that is, low rate after reinforcement with positive acceleration up to next reinforcement
Reinforcement after varying intervals with a mean interval specified	High stable rates, very great resistance to extinction
At least a minimum interval between responses to be reinforced. Responding prior to the interval resets the clock timing the interval to zero	Stable low rates
Rate of response must be at a specified high value for reinforcement to be given	Stable high rate
Two or more consecutive simple schedules; S^d indicates each; primary reinforcement after all component schedule requirements have been completed	Varies with variations in component requirements. Components influence each other. The last component is especially influential, presumably because the frequency of reinforcement in that component controls the reinforcing properties of its S^ds, which in turn, controls behavior in preceding components
Two or more consecutive simple schedules; no S^ds; primary reinforcement after all component schedule requirements have been completed	Depends on component schedules, size of ratio or interval of components, and sometimes on the order of components
Two or more simple schedules; each indicated by S^d; primary reinforcement after completion of each component (called *ply*)	Behavior in component plys resembles that of simple schedule alone. Sometimes interactions occur, such as contrast effects (in which rate of responding in one ply is inversely related to frequency of reinforcement in a neighboring ply. Interactions are controlled by briefly withdrawing the opportunity to respond before shifting to a new ply (called "Time out")
Two or more simple schedules sequentially presented; not distinguished by S^ds; primary reinforcement after each component	Behavior not strongly under control of individual schedules, more under combined control of simple schedules, as when two FRs are combined for a VR. Rates higher than on multiple schedule
Two or more simple schedules simultaneously	Rates tend to match frequency of reinforcement in a given option.

tendency to change the response strengths that were there at the end of training.

During training the experimenters are establishing a baseline. They would like to maintain that exact baseline throughout post-training testing. But this is not possible. If they reinforce behavior during testing, they are increasing response strengths to the stimuli; if they do not; the baseline will be lost as a result of extinction.

Stimulus generalization experiments require a control for baseline changes during testing. Most experiments on generalization involve testing without reinforcement ("under extinction"). It is essential not to confound the resulting loss of response strengths with the influences of stimulus similarity that are of prime interest. How can we control for that?

It is like controlling for any other order effect. If we cannot eliminate a variable, we can at least extract its effect *statistically* if we randomize or counterbalance order of presentation. If the test stimuli are presented in varying orders, the influence of extinction will be equally represented for each of the test stimuli. In this way, the influence of extinction is *held constant.*

To illustrate, in a classic experiment on the gradient of stimulus generalization, Hovland (1937) measured generalization by using a within-subjects design with grouped data. He used four tones, and any given subject was trained to give a conditioned response to only one of the tones, but was later tested for response to all four. One group of subjects was trained on the highest tone and tested on the other three, and another group was trained on the lowest tone and tested on the other three. These data were ultimately grouped, thus counterbalancing for the effects of using a high or a low tone for CS. The subjects in this experiment were humans, and the response was a galvanic skin response, a change in the electrical resistance of the skin, which Hovland induced by giving subjects an electric shock as unconditioned stimulus. During generalization testing, no UCS was presented; in other words, generalization testing took place during extinction. In a design of this sort, which employs the comparison technique, the effects of extinction are important and might obscure the actual gradient of generalization unless special precautions are taken. Accordingly, Hovland presented the test stimuli to different subjects in different orders (counterbalanced order of presentation) so that the effects of extinction would be equal for each stimulus. The resulting gradient is presented in Figure 6.3, in which the results show that the magnitude (degree of generalization) of the GSR decreases progressively with increased differences in pitch, producing a gradient that is curved downward.

To a degree, *experimental* control can be exerted over effects of extinction. Even though we cannot eliminate these effects entirely, we can reduce them to a very low level by taking advantage of what we

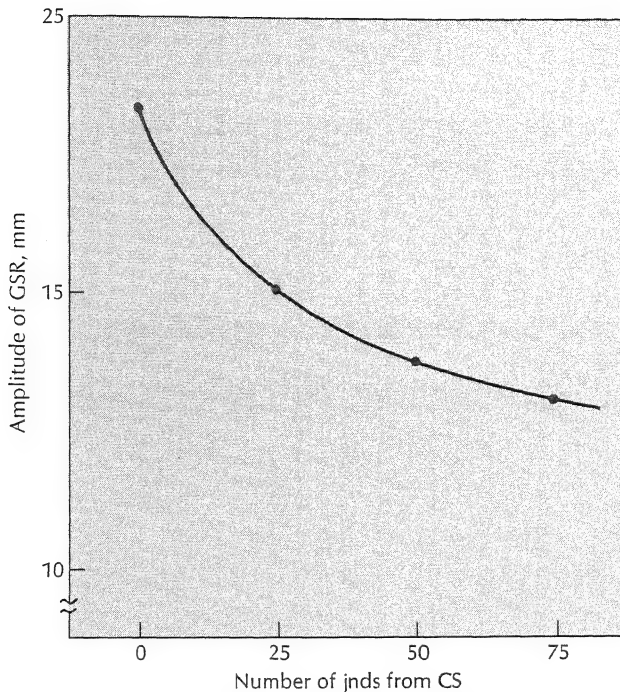



FIGURE 6.3 Stimulus generalization gradient for the galvanic skin response conditioned to a tone of 1000 H_z. (From Hovland, 1937.)

know about intermittent reinforcement. Many investigators have used variable-interval schedules of reinforcement while establishing the baseline level of stimulus control. The great resistance to extinction produced by such schedules reduces the influence of extinction to a minimum. Still, it is necessary to randomize or counterbalance order of stimulus presentation during testing, since the order effect due to extinction has not been eliminated completely. An experiment by Guttman and Kalish (1956) illustrates the value of an intermittent schedule in obtaining a gradient of generalization.

The Hovland (1937) technique entails the use of grouped or averaged data. It has often been pointed out, as an argument against the acceptance of a generalization gradient of the form obtained by Hovland, that no individual in Hovland's experiment showed a gradient. This is a trivial objection because, for any given individual, the effects of extinction are confounded with the generalization effects. It is only with averaged data that the effects of extinction can be said to be equally distributed among the stimuli.

Grouped data do not always reflect the functions that are actually present in the individuals composing the group; experimenters would prefer to observe a generalization gradient in individual organisms. Guttman and Kalish (1956) devised a technique that yields individual

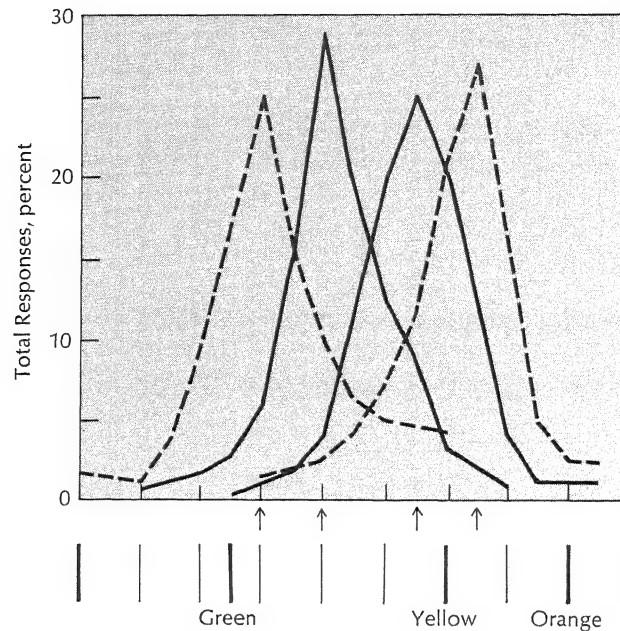



FIGURE 6.4 Generalization gradients of pigeons along a wavelength dimension with light as the stimulus. The arrows indicate the stimulus that was present during variable interval reinforcement. Response rates diminished with distance from the original discriminative stimulus. (From Guttman, N., & Kalish, H. I., 1958.)

gradients. Training pigeons to make an operant key-pecking response on a variable-interval (VI) schedule, they got a very stable baseline.

The key was lighted from behind with a monochromatic light. For a light to be monochromatic it must be purely of one color (one wavelength). The VI schedule gives rise to very great resistance to extinction. So it was possible to present a variety of stimuli to a subject before the effects of extinction were marked. Generalization testing therefore took place without reinforcement. The key was illuminated by lights of various hues at different times during the test session. The wavelengths of the lights included the one present during maintenance of the VI schedule and those both above and below it. The frequency of responding in the presence of these various hues is taken as the measure of generalization. The kind of gradient that resulted is illustrated in Figure 6.4.

Even though various pigeons respond at individual over-all rates, the *form* of their gradients is remarkably similar. It is much like that obtained by Hovland (1937). However, the response rates in the Guttman and Kalish study tend to be slightly higher for stimuli very near the stimulus used during training.

If an experimenter could find a way to measure gradients of stimulus

generalization without any extinction at all, it would be ideal. Of course, this would have to be accomplished without introducing some new interfering variable. Even better than an experiment to give gradients without extinction would be an experiment yielding gradients with *and* without extinction, so we could see if there is any influence of extinction on the form of the gradient. The Guttman and Kalish (1956) procedure comes close, but even a VI schedule permits *some* extinction.

Wickens, Schroder, and Snide (1954) devised an experiment that gave a stimulus-generalization gradient on the very first test trial, prior to any extinction. They borrowed many of Hovland's (1937) procedures, even using three of the tone frequencies used by Hovland. They changed Hovland's procedure by using only *one* test tone for each subject, instead of giving each subject *every* tone during testing. Using a large number of subjects, they plotted separate generalization gradients for single test trials, such as trial 1. Theirs was a mixed factorial design with independent subjects receiving different tones and each subject receiving repeated trials. Figure 6.5 shows the gradients for trials 1, 2, 3, and 8.

Notice that there really is not much of a gradient for the early trials. Responses are about as great to the test stimuli as they are to the stimulus used as CS (the one labeled as having "zero difference" on the graph). Such gradients are said to be "flat," and are usually taken to mean that the behavior is not under stimulus control.

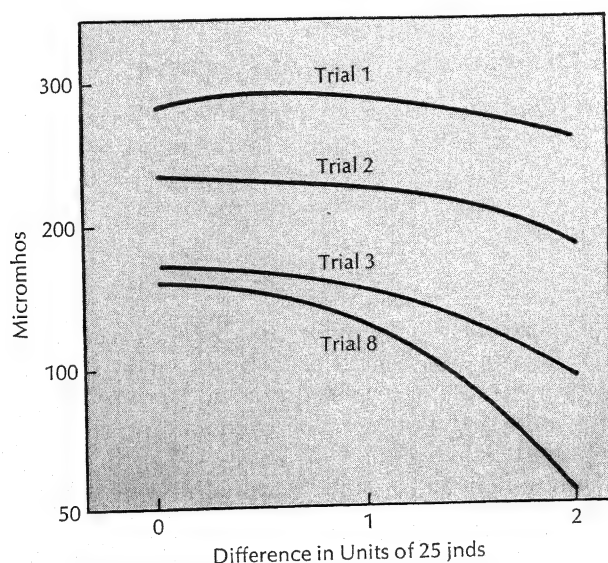



FIGURE 6.5 Smoothed generalization gradients at various stages in extinction. (From Wickens, Schroder, & Snide, 1954.)

Key Ideas Box 6.4: Stimulus generalization

After the formation of a stimulus-response connection, similar stimuli will control the response. This is the *principle of stimulus generalization.* The notion of a *gradient of stimulus generalization* is that the degree of responding varies as a function of the degree of similarity to the conditioned stimulus.

It is difficult to measure stimulus generalization unobtrusively. While testing for tendencies to respond to similar stimuli, we must either continue reinforcing the subject or not (extinction). Both procedures change the baseline behavior we want to measure. Usually testing is done under extinction. This produces an order effect that must be controlled. Randomization or, more often, counterbalancing of order of presenting the stimuli is generally used when designs involve, as they usually do, repeated measures on the same subjects. Intermittent reinforcement is often used to minimize the effects of progressive extinction. It is still necessary to randomize or counterbalance. An alternative is to test each subject on only one stimulus, shifting to a completely randomized design. Unfortunately, the different procedures tend to produce slightly different gradients of generalization.

General and specific laws of learning

It is clearly *legitimate* to emphasize the search for highly general laws of learning. However, many investigators have come to believe that it may not be the *best* strategem. Evidence has accumulated that there are very important characteristics of learning that are specific to given types of organisms.

Specific hungers

Many organisms can select foods that contain substances for which they have a deficiency. Sometimes, as in the case of deficiency for sodium, the preference requires no learning at all! Commonly organisms immediately prefer foods containing sodium if the diet has been sodium-deficient (Rozin & Kalat, 1971). In other cases, for example, thiamine deficiency, some learning may be required before they make appropriate selections. Rozin and Kalat (1971) point out that rats mainly learn an aversion for the deficient food. This occurs despite the large time gap between dietary deficiency and the development of ill

health. Of all the possible stimuli that occur prior to sickness, rats select the *diet* as the one to which they develop an aversion. They are not averse even to such things as the familiar food dish.

The selection of a particular stimulus to avoid seems to be specific to given species. A similar case has been reported of taste aversions in rats learned from associating tastes or odors with malaise induced by X rays or apomorphine (Garcia & Ervin, 1968). There is a special readiness to associate specific stimuli with specific responses. Wilcoxon et al. (1971) have shown that whereas rats tend to associate induced illness with taste cues, quail associate it with visual cues. There seems to be a preparedness of the nervous system to develop certain kinds of conditioned response, and the preparedness depends on adaptations that make evolutionary sense (Rozin & Kalat, 1971).

A related phenomenon is called "instinctive drift." It was described by Breland and Breland (1961) in a classic paper on the "misbehavior" of organisms. While trying to develop operant responses in a variety of species, they found them sometimes quite resistant to the contingencies of reinforcement. For example, raccoons tend to wash their food before eating it, and they will persist in doing so even if the food is sugar that dissolves when washed!

Seligman (1970) argues that organisms may be *prepared, unprepared,* or *counterprepared* to associate certain events. He cites many experimental examples, and speculates that the laws of learning may vary with the preparedness of the organism for the association.

Key Ideas Box 6.5: General and specific laws of learning

If we seek the most general laws of learning, we can study any convenient species of subject and any arbitrarily selected response. Psychologists have assumed that there are such laws and that they should be studied first. However, recent work has indicated that different species and different types of responses may follow somewhat different laws. Organisms seem to be *prepared* to acquire certain responses or stimulus-response connections. They may also be *unprepared* or even *counterprepared* to learn certain things. This suggests that, though it is legitimate to study arbitrarily selected responses of arbitrarily selected species, a better strategy might be to take into account specific characteristics of species and their responses. The latter point of view may entail a strategic decision that will profoundly transform methods of studying learning.

Autoshaping

Thus, the notion that responses, stimuli, and organisms may be selected arbitrarily in order to find general laws of learning is under fire. Even the most widely used response in operant conditioning, the key-peck response of the pigeon, seems to have features not clearly related to contingencies of reinforcement. Brown and Jenkins (1968) showed that pigeons will peck a lighted key when the light and the reinforcer are paired without regard to the pigeon's behavior. This is called *autoshaping.* Pigeons seem to have a tendency to peck at stimuli that are predictive of reward (Bilbrey & Winokur, 1973), even if reinforcers are not contingent on the key pecking.

The tendency to peck at a lighted key persists even when an "omission contingency" prohibits delivery of food on trials during which pecks occur. The name given to this phenomenon is *automaintenance.* With the omission contingency, responses would appear to be discouraged by the contingencies of reinforcement. However, Lucas (1975) showed that the key pecks are maintained by the reinforcement of pecking movements made *toward* the key without actually touching it. That is, a pigeon may make a pecking gesture toward the key without actually touching it. Even if the experimenter has arranged to prevent rewards for pecking, these gestures may be reinforced, thereby maintaining the class of behaviors of which pecking is a member.

Study Questions

1. What is vicious circle behavior?
2. What is errorless discrimination learning?
3. What is fading?
4. What is aversive control and what are its main types?
5. What is respondent conditioning?
6. Present and explain Pavlovian conditioning.
7. Explain the following terms:
   a. acquisition
   b. latency
   c. asymptote
8. Give experimental evidence of the phylogenetic generality of respondent conditioning.
9. What kinds of stimuli can function as a neutral stimulus in respondent conditioning?
10. Discuss the generality of respondent conditioning across responses.
11. Distinguish between operant and respondent conditioning.
12. Distinguish between a controlled and free operant situation.

13. What are:
    a. operandum
    b. manipulandum
    c. magazine
14. Explain the following:
    a. operant level
    b. magazine training
    c. shaping
    d. successive approximations
15. Discuss the phylogenetic generality of operant conditioning.
16. Explain the following:
    a. extinction
    b. criterion of extinction
    c. disinhibition
    d. spontaneous recovery
17. What is a discriminative stimulus?
18. What do the following symbols mean:
    a. S^d
    b. S^{Δ}
    c. S^+
    d. S^-
19. What is the difference between a procedural and a functional definition of operant conditioning?
20. What is respondent–operant overlap?
21. Distinguish between punishment and negative reinforcement.
22. Distinguish between positive, negative, primary, and secondary reinforcers.
23. What is continuous reinforcement and what kind of behavior does it produce during maintenance and during subsequent extinction?
24. What is a variable-ratio schedule, and what behavior does it produce?
25. What is a fixed-interval schedule?
26. What is a variable-interval schedule, and what pattern of responding is typical of it?
27. What data have led certain theorists to oppose the idea that it is an optimal strategy to look for general laws of learning by studying arbitrary responses?
28. Distinguish between the gradient of stimulus generalization and the principle of stimulus generalization.

7 LEARNING
Selected Research

Insight versus trial-and-error learning

A good deal of research in the psychology of learning has been done on animals, especially rats. And two major theoretical views underlying much of this research have been the *cognitive* and *stimulus-response* viewpoints.

Stimulus-response (S-R) theorists, much concerned to save psychology from slipping back into its age-old unscientific ways, tried to account for behavior in the simplest mechanistic terms. They viewed conditioning as mere strengthening and weakening of bonds between stimuli and responses, and they regarded more complex learning as an edifice whose brick and mortar consists of stimuli, responses, and the bonds between them.

Cognitive theorists, on the other hand, insisted that learning has facets other than the mere formation of bonds between stimuli and responses. They argued, for example, that organisms, even simple ones, learn the layout of things, learn to expect what leads to what. These expectancies have been called "cognitive maps." Cognitive theorists have been referred to as S-S (stimulus-stimulus) theorists. The S-S notion seemed to provide a nice symbolization of the notion of expectancy.

The struggle (and an intensely emotional struggle it was) between cognitive and S-R theorists provided a source of energy fueling a great proportion of the research in animal learning that took place over the last twenty years or so. One could hardly make sense of the main thrust of the experimental psychology of learning without having some concept of that struggle.

I have selected a particularly influential dispute as a sample. It is the dispute over whether learning occurs by insight or by trial-and-error. Cognitive theorists cast their lot in favor of insight, and S-R theorists fought them tooth and nail.

Not only does this issue provide a good example of a central tension that impelled students of learning to action; it also illustrates several methodological points of quite general interest to the experimental psychologist. It is a general rule that scientific questions proceed from the vague to the specific, not from the particular to the general (as is commonly supposed). This may be seen here as the development of the crude issue of insight versus trial-and-error into increasingly precise questions. To watch and understand this development is to see why creative and intelligent psychologists are willing to spend their lives working on research problems that seem from the outside to be mere trivia.

It is not easy to distinguish the increasing precision of mature science from its counterfeit, the misguided pursuit of the minute and irrelevant. We will likely have to concern ourselves with small details of apparatus and procedure before we have a deep grasp of psychology. And many of the questions we work to answer may well seem paltry to those who stand outside and view our work from a distance, without seeing its meaning.

Scientific progress often lies in the crispening of questions. We get the impression that scientists begin with clear-cut questions, put them in the form of hypotheses of "if . . . then" statements, test the implications empirically, and write the results for publication. This description of the scientific process is not so much wrong as it is incomplete. Or perhaps we should say that it has the wrong emphasis.

Scientists often do go through the formal hypothesis-testing process. However, they often end up not with an answer to the original question but with a better understanding of the right kinds of questions to ask. This is one reason why it is a good idea to jump right in and start doing research, all the time keeping your wits about you and thinking hard about what you are doing and why. It is unlikely that you will discover the right questions without doing research first.

This and several other tactical points are illustrated by the insight versus trial-and-error issue. The dependence of functional relations on techniques of data analysis, the role of the type of apparatus used in determining the experimenter's interpretation of the psychological

process under investigation, and the dangers of nonrigorous reasoning are all brought home within the framework of this question. Take care, then, to look beyond the particulars of the discussion while keeping in mind the general implications of the tactical problems encountered.

THORNDIKE'S EXPERIMENTS

A classic series of experiments done by E. L. Thorndike indicated that learning occurs gradually by trial and error or by trial and success (Thorndike, 1898). Thorndike's early work was done on a variety of animals, including fish, chicks, cats, dogs, and monkeys. Later, he extended the work to human subjects and made very important contributions to educational psychology. One of his best-known experiments, and a typical one, was done on cats in puzzle boxes. A typical puzzle box is illustrated in Figure 7.1.

The task of the cat in the puzzle box was to discover a particular response such as pulling a string or depressing a lever that caused the box to open and allowed the cat to escape to freedom and some sort of reward.

Figure 7.2 shows a typical performance for a cat in a puzzle box. The cat began by making a variety of movements, usually described as "random." After this period of poorly directed movement, the animal made the correct response and was able to escape. The time required for this first escape is represented by the first point in Figure 7.2. Upon the cat's being replaced in the apparatus, one might expect that it would immediately make the correct response. However, the learning curve shows that, though the correct response was made sooner on trial

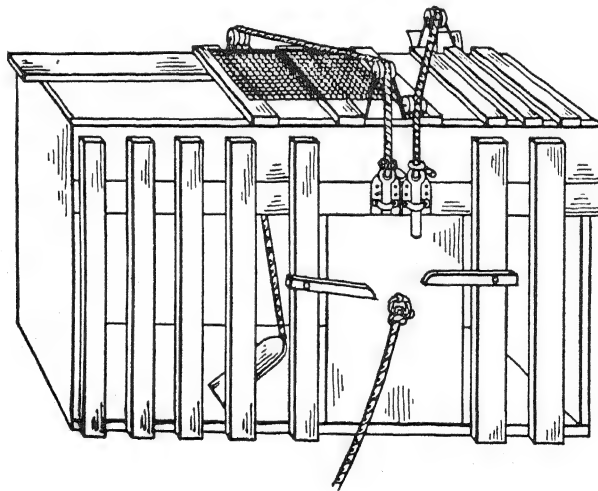



FIGURE 7.1 A typical puzzle box used by Thorndike in his classic studies of animal problem solving. (From Garrett, 1951.)

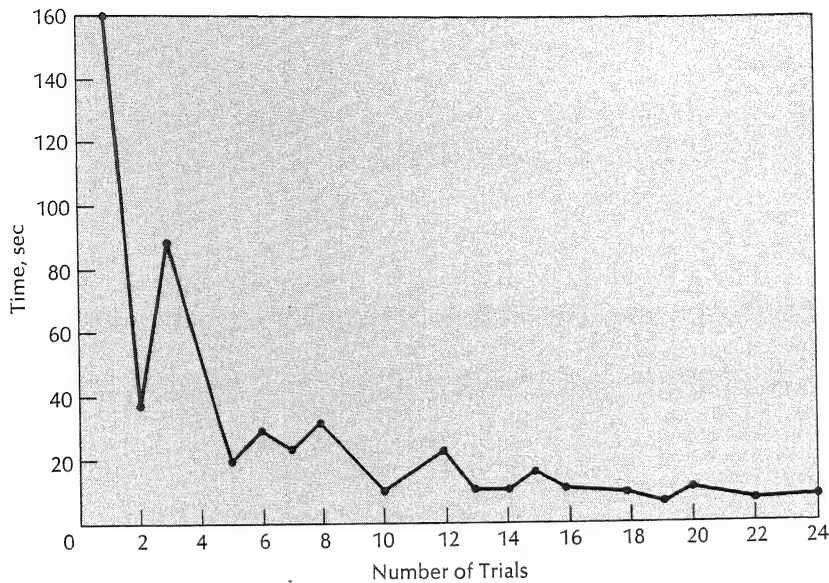



FIGURE 7.2 A typical performance curve for a cat in a puzzle box. (From Garrett, 1951.)

2 than on trial 1, it was by no means made immediately. In fact, there was a gradual reduction in the amount of time required to escape, until finally the animal achieved a stable performance of rapid problem solution. At no time did the cat show any sudden flash of insight. This would be revealed in a marked dip in the learning curve, followed immediately by stable performance. In general, then, the studies of Thorndike lead us to suppose that the learning process is gradual (trial-and-error) and not the result of insight.

KÖHLER'S STUDY OF APES

During World War I, the gestalt psychologist Wolfgang Köhler conducted experiments that led to a very different conclusion from that of Thorndike. He spent a number of years on an island that had a colony of captive chimpanzees for study and observed many aspects of their behavior. Köhler (1925) reported his observations in his now famous book, *The Mentality of Apes.* Time after time he reported cases of what appeared to be insightful solution of problems. Typically, the animals made futile attempts to solve a problem, gave up for a time, encountered something that made solution possible, and immediately dashed back to the problem to apply the newly found solution.

Probably the most widely cited case is that of the chimpanzee named

Sultan. He was faced with the task of combining two short sticks into one long one in order to reach a food reward. Sultan at first made a variety of futile attempts to reach his food by jumping, using one stick, and so forth. Finally the animal gave up and went away to sulk. While sulking, Sultan idly manipulated the sticks. At one point in the manipulation the two sticks came together in a way that permitted them to function as one long stick. Sultan immediately got up, went over to the food, and retrieved it with the newly discovered tool.

This performance contrasts rather sharply with that of Thorndike's trial-and-error learners. Not only was the chimpanzee able to apply the solution immediately but, unlike Thorndike's subjects, he applied the solution with all the vigor of certainty despite his never having tried it before. Remember, Thorndike's cats even failed at immediate application of the correct solution after having been successful at using that solution several times.

SITUATIONAL SPECIFICITY

Why do the experiments of Thorndike and Köhler give us such different notions of the learning process? Could the differences be due to the use of anthropoid apes in the one case and lower animals in the other? No, because Thorndike did some work with primates and Köhler saw evidence of insightful learning by lower animals such as dogs.

Köhler (1925) suggested an alternative account of the discrepancies, and this account has broad implications for all branches of psychology. It is simply this: The apparatus and procedure selected by experimenters may largely dictate the kind of result they will get. In other words, functional relationships are to some degree situationally specific or relative to the conditions under which observations are made. This is a very general principle.

Köhler argued that the kind of apparatus used by Thorndike could not *in principle* reveal insight even if the animal subjects possessed the capacity to reveal it under proper circumstances. From Köhler's point of view, some training situations will and others will not result in insightful performance. Those that will not, such as Thorndike's mazes and problem boxes, are simply the wrong kinds of apparatus. In the wrong apparatus it is impossible for the organism to see the relevant aspects of the situation as a whole, so it cannot possibly generate a solution based on insight into the essence of the problem.

May we conclude that learning can occur by insight without benefit of previous trial and error? Not really. Köhler did not have a clear conception of the amount and type of experience his chimpanzees had. It would be possible for them to have gone through some relevant trial-and-error learning before they came into his possession. This

Key Ideas Box 7.1: Insight versus trial-and-error learning: Early work of Thorndike and Köhler

Does learning occur by sudden insight, yielding a curve that jumps abruptly from chance to high-level performance?

Thorndike studied animal learning and found that learning was gradual. He maintained that learning occurs by the gradual stamping in of correct response tendencies and the gradual stamping out of erroneous tendencies. Hence, he supported the notion of trial-and-error learning, in which organisms are said to respond "randomly" at first and then, through gradual strengthening of appropriate response tendencies and weakening of inappropriate ones, to accumulate a basis for high-level performance. In a typical experiment, a cat was placed in a puzzle box from which it could, by pulling a rope suspended in the box, escape and gain a reward. The amount of time taken to escape declined *gradually* with repeated trials.

Köhler observed chimpanzees in a variety of problem-solving situations, and noted sudden, insightful learning. For example, one ape was faced with the problem of getting at suspended bananas that were too far up to be reached by jumping, even with the aid of one of the sticks placed in the cage. The sticks fitted together, however, in a manner that would permit them, in combination, to reach the reward. All early solutions having failed, the ape dawdled with the sticks, and happened to stick them together. It immediately ran over and used the newfound device to retrieve the bananas.

Köhler argued that animals *do* learn by insight. He maintained that apparatus of the type used by Thorndike failed to tap the real capacities of organisms, since their arbitrary nature made it impossible to perceive a meaningful relationship between the problem and its solution.

Later experimenters (such as Schiller) showed that primates will only show insightful solutions of problems like the ones Köhler used after they have had a good deal of previous experience playing with sticks. Modern studies of chimpanzees in the wild suggest that they commonly go through a process of learning to use "tools."

conjecture is supported by the report of Birch (1945), who studied chimpanzees that had been born in captivity and raised under controlled nursery conditions. He investigated their stick-using behavior and concluded that the perception of such relationships depends on the previous experiences of the animal. He maintained that insightful

problem solution is due to the integration into new patterns of previously learned component responses.

Schiller (1952) also studied manipulation of sticks by chimpanzees and found that it occurred independently of externally applied rewards. For example, when two short sticks that could be fitted together were given to each of 48 chimpanzees, 31 of them fitted the sticks together within 1 hr. This means that there was every reason to suppose that Köhler's subjects might have had prior experience relevant to the tasks he gave them. It is worthwhile to note also that Goodall (1968) has reported that chimpanzees in the wild commonly construct and use simple tools.

The continuity-noncontinuity controversy

The continuity-noncontinuity controversy is a refinement of the insight versus trial-and-error issue. The question of insight versus trial-and-error was too broad for concise investigation. The continuity-noncontinuity controversy led to increasingly precise formulations of the meaning of the question and to increasingly subtle experimental tactics. The component questions become more and more clearly defined. They also became more and more remote from the questions of everyday life. Science inevitably progresses in this way. It is hard to find an instance of any question being answered scientifically without its first having been refined to a point where the ordinary person cannot, without special instruction, see its relevance. That is why scientists seem so intellectual and their work so complicated. We come into the scientist's dialogue with Nature when it is very late in the conversation.

The continuity-noncontinuity controversy was launched by I. Krechevsky (now known as David Krech). He did a very intelligent thing while studying rats as they learned in discrimination boxes. Prior investigators had assumed that when rats responded at chance levels during early trials they were responding "randomly." Krechevsky realized that their behavior was random only with respect to the correct cue as he, the experimenter, had defined it. He decided to look at the discrimination from various other points of view, perhaps even the rats. He found that the rats were responding quite systematically with respect to cues other than the one the experimenter had in mind.

To be specific, suppose a rat is learning to discriminate a black from a white cue.[1] The experimenter might well find that the rat will go to the two cues about equally often. Gradually the rat will begin to go

[1]Generally experimenters counterbalance the cues so that half of the subjects get one cue as correct and half get the other. Otherwise the percentage of total "correct" choices before shaping might not approximate the chance level of 50 percent. For example, rats often prefer black to white.

more and more often to the cue that leads to reward. Finally it will do so often enough that it will reach the experimenter's operational criterion of "learning," say 90 percent correct for 10 trials in a row. This surely looks like random responding followed by gradual strengthening of the correct response.

If you now look at which side the rat goes to in the early trials, you will usually find that it is not random at all. Most rats have a *position preference*—they go to one side with great consistency. So with respect to the position cue they are not behaving randomly. In fact, they behave quite systematically with respect to one cue after another, until they finally start responding as the experimenter intended them to do. Early learning is, therefore, not random, but systematic. Krechevsky took the point of view that the rats systematically tested "hypotheses" in the early trials. They tried this cue, then that, until they fell upon the correct one. Once they had picked the right *class* of cue, learning to go to the correct member of that class was easy. Thus, they have to learn to stop paying attention to which side they go to (the spatial class of cue), which smell they go to (the olfactory class of cues), or any other cue dimension except the dimension of "brightness" that the experimenter has in mind.

Notice that from Krechevsky's point of view there are two stages to the learning process. One stage involves learning to attend to the correct class of cue. The other involves learning to approach the particular member of that class that has been defined as correct. The second stage takes relatively little time, so once the animal has discovered the right type of cue, its learning is very rapid. For this reason, Krechevsky's point of view is a *noncontinuity* one. You can see that, like Köhler, Krechevsky would expect sudden solutions to problems. He would expect discontinuities (or noncontinuities) in the learning process.

Krechevsky's work, like that of Köhler, illustrates the importance of the way you measure behavior as a determinant of how the behavior seems. Measured in the manner of earlier investigators, early learning seemed random. By changing the method of measurement to that of Krechevsky, we can see systematic patterns. It is very important to be flexible about methods of measurement, because a slight change of measure can bring about radical transformations of the interpretation of phenomena.

There is a difference between Krechevsky's experimental findings and his interpretation of them. It is a *fact* that rats behave systematically in early trials. Regardless of the interpretation, Krechevsky gets credit for finding a way to measure that. He interpreted the fact to mean that rats test hypotheses. He was not entirely clear on just what that meant. But the essence of it seems to have been that they sampled one cue dimension after another until they fell upon the rewarded one.

Key Ideas Box 7.2: Continuity versus noncontinuity learning

The "insight versus trial-and-error" issue developed into two more modern controversies; the "continuity-noncontinuity" issue and the "all-or-none versus incremental learning" issue. The continuity-noncontinuity controversy centered on the manner in which animal *discrimination learning* occurs. *Continuity theorists* regarded discrimination learning as a combination of conditioning and extinction. Approach tendencies to the available stimuli changed as a result of selective reward and nonreward for approaching them. The animals' discriminative responding was seen as a function of the relative approach tendencies thereby accumulated. The continuity theorists thus attributed discrimination performances to a single factor, the relative strengths of approach tendencies to stimuli (variously called "habit strengths," "response strengths," "associative strengths," or something of that sort).

Noncontinuity theorists argued that there are *two* factors operating in discrimination learning. The first is a *"stimulus selection"* or *"attention"* factor. An animal was not behaving randomly in early trials. It was selectively attending first to one, then to another property of the stimulus, while ignoring all others. Thus, it might begin by attending to the positions of the stimulus, ignoring such things as brightness, odor, or pattern. Next it might attend to odor, next to brightness, and finally to pattern. If the experimenter identified pattern as the correct dimension, the subject could now turn to the second stage of learning—the discovery of which pattern had to be approached in order to get rewarded. This second, stimulus-response approach tendency is like that of the continuity theorists. Where the two theories differed was on the point of whether the organism actively selected the stimulus dimension that would control its behavior.

Thus, at its core, noncontinuity theory says that discrimination learning takes place in two stages. The organism must first learn to attend to the correct type of cue, and then must learn to approach the one of that type that is correct.

ENTER KENNETH SPENCE

The continuity side of the controversy was represented by Kenneth Spence. He felt very strongly that discrimination learning required no

two-stage model for its explanation. Note that he did not disagree with Krechevsky's facts. That would be rare for a scientist. He disagreed with the interpretation of those facts. He argued that the systematic patterns of responding would be produced even if learning has only one stage, that of connecting the correct response to the correct cue. No stage of searching for the correct dimension was necessary.

Spence devised a simple mathematical model that assumed that learning is merely a matter of combined conditioning and extinction, with no "hypothesis testing." He showed that such a model would produce the early systematic behavior Krechevsky had observed.

This provides us with an important lesson on the relationship between theory and fact. Sometimes it is not obvious which facts a theory will predict until the implications of the theory have been worked out in detail. The result of this is that scientists have to be very meticulous in working out their theories. And it therefore takes time and patience to appreciate the theories. I will not try your patience by working out Spence's theory in detail. You can either take my word that it works or look it up in Spence (1936). If you are at all theoretically inclined, it would be worth your while to look it up.

Behavioristic psychologists, of which Spence was a prime example, are most unlikely to settle for any theoretical account. They like data. So Spence went on to dedicate the work of his laboratory to experimental tests of the noncontinuity versus the continuity theory. You see, his theory had left matters up in the air. Both his and Krechevsky's theory seemed able to account for Krechevsky's data. Many investigators devoted their energy to resolving the issue. It is hard for us today to imagine the energy, enthusiasm and even anger that went into this controversy. The flavor of that is something I will not be able to convey. Let us content ourselves with a description of the major tactics devised and implemented in order to resolve the issue.

Presolution reversal of cues

The two theories seem to make different predictions concerning the effect of interchanging the correct and incorrect cues during early trials. Suppose you use a completely randomized design with two groups. One group of rats is trained on a discrimination of black from white. The other group is trained the same way during the early trials, but whichever cue was correct (say black) at first is later made incorrect. The formerly incorrect cue is, of course made correct.

What would noncontinuity theory predict about the relative numbers of trials taken by the two groups to reach a criterion of learning? Assuming that the shift is made while the animals are still responding at 50 percent correct, the reversal of cues should have no effect. The noncontinuity viewpoint would be that the animals are paying attention to something else during the presolution phase. If you switch the

brightness cues around while they are attending to spatial cues, they should not even notice. Thus, noncontinuity theory predicts no reliable difference between the two groups.

In contrast, the continuity theory predicts that the group receiving the reversal will take more trials to reach criterion. They will build up a conditioned tendency to approach the originally correct cue. When the reversal occurs, the earlier approach tendency will have to be broken down before the correct tendency can be built up. Notice that an outcome in favor of a noncontinuity position requires acceptance of a null statement. This tactic is capable of giving strong support to the continuity position, but only weak support to a noncontinuity position. If there is a difference between groups, the continuity position is firmly supported. If there is no difference, the noncontinuity position is supported, but only weakly, by a null acceptance.

Actual results from studies using the tactic of presolution reversal of cues yielded varying results. Taken as a group, the studies could not make a case for either point of view, though individual studies seemed to support one or the other. For example, McCulloch and Pratt (1934) trained rats on a string-pulling response in which differing amounts of weight tied to the strings were to be discriminated. One group was run directly to criterion. A second group underwent reversal of cues during the presolution period. The second group learned much more slowly, supporting the prediction of continuity theory.

On the other hand, Krechevsky (1938) trained three groups of rats in a Lashley jumping stand (see Figure 2.3) to discriminate horizontal from vertical rows of small black squares. One group was run directly to criterion, a second encountered reversal at the twentieth trial, and a third underwent reversal at the fortieth trial, which is close to the time when learning begins to reveal itself. The group reversed at the twentieth trial was not slowed in learning, thus supporting the noncontinuity viewpoint. The 40-trial group was slow to learn, however. How did Krechevsky deal with that? He interpreted it to mean that the rats were beyond the presolution phase by the fortieth trial. That interpretation placed the noncontinuity theory in serious jeopardy.

We said earlier that the facts of science are characterized by reliable observation and the theories must make risky predictions. Theorists who change their theories, claiming that unexpected facts fit in with them all along, are not scientific theorists. Krechevsky had earlier indicated that the presolution phase was identifiable as a period of responding at chance levels with respect to the experimenter's cues. But the group receiving reversal at 40 trials was not responding at chance levels. He still wanted to consider the first 40 trials part of the presolution period. Not only did this call into question the ability of his theory to make risky predictions; it also caused doubt as to whether

the very concept of "presolution period" could be measured. Its standard operational definition was being abandoned.

Spence did not hesitate to point these things out to Krechevsky (Spence, 1940). But whatever errors Krechevsky may have made, the group receiving reversal at 20 trials posed a problem for the continuity view. How could Spence account for its behavior? He argued that the animals had not yet learned to look at the spot where the relevant cues were located. His argument was that rats look where they are leaping. Krechevsky's cues were apparently placed somewhat above that.

Scientists argue a great deal, but at a certain point they go to the data. Ehrenfreund (1948) did a study in which cues were located either where the rats hit the cards or above that point. The behavior of those with the cues at the jumping point supported the continuity view, whereas the other group's behavior was consistent with noncontinuity, as Spence had predicted.

At this point you might well think that the case for the continuity viewpoint is beginning to get fairly strong. But there is an important flaw in Spence's argument. It is so important that it might even mean that Spence inadvertently slipped into a noncontinuity viewpoint. It is simply this: a rat's visual field encompasses over 300 degrees (Lashley, 1942). Rats can see all around themselves but for a wedge of about 60 degrees directly behind them. How, then, could they fail to see cues a few centimeters above where they were jumping, unless by *directing their attention*? This is critical, because the directing of attention is exactly what noncontinuity theory was trying to reveal. Continuity theorists talked as though it was a mere matter of whether or not the rays of light from the stimuli reached the eyes. But it had to be a matter of selective attention.

The end product of research with the tactic of presolution reversal of cues was more a sharpening of the wits, an improvement in understanding of how intricate an experiment must be to solve the problem, than a decision as to which theory was correct. It is time we moved on to another tactic.

The tactic of postsolution cue-substitution

Early in the argument, Spence attributed a noncontinuity point of view to Karl Lashley. Lashley did not feel identified with that point of view, but soon gained an interest in the issue and decided to take up the point of view Spence had attributed to him. He devised the tactic of postsolution cue-substitution. The noncontinuity view is that organisms learn to attend selectively to a particular kind of cue. This means that, once they have learned a discrimination, they should be attending to the experimenter's type of cue and ignoring other types. If at that time an extra cue were put in, one of a different type, that was also

consistently rewarded, what would happen? Remember, the new type of cue is *extra.* The animal can still go on using the cue it learned about in the first place.

From a noncontinuity point of view, the animal should learn nothing about the new cue. Attention should be riveted on the old type of cue. From a continuity point of view, there is no selective attention. Conditioned approach tendencies should build up to the new cue. Thus a great deal should be learned about the new cue if continuity theory is right. Nothing should be learned about it if noncontinuity theory is right.

Lashley (1942) put it to the test. He trained rats in a jumping stand with pattern size as the relevant cue. Large and small circles were used. (Let's say the large circle was the correct cue). After they had learned, he replaced the large *circle* with a large *triangle.* Now the rats could go on performing well by using the old cue of size, but shape was an extra cue. After much training, he tested the rats by eliminating the cue of size. He presented them with a triangle and a circle of about the same size. They showed no preference. Thus, they seemed not to have noticed the extra cue, as noncontinuity predicted they would not.

Unfortunately, in this case we have a problem of replicability. A later study used the same tactic and got the opposite results (Blum & Blum, 1949). The study of Blum and Blum was a systematic, not a direct, replication of Lashley's. Thus the problem may have been one of generality rather than of reliability. However, the predicament we face in interpreting such contrary findings brings into focus a central strategic problem. Were the methods of establishing reliability and generality adequate? What should they have done?

There are certainly no generally accepted answers to these questions. Indeed, their methods differed little from those of many investigators today. I argued in Chapter 2 that there has not been enough emphasis on establishing the generality of findings prior to publication. An assessment of generality might well have revealed to Lashley that there were critical variables that had to be at the right levels for his result to be replicated.

The emphasis on generality is still not as great as would be ideal, but there has been an important improvement since Lashley's time. Today, investigators are much more likely to conduct experiments in which several variables are manipulated (multifactor designs). Thus a partial assessment of generality is incorporated into their studies.

More could be said about this, but let's look at other tactics used to resolve the controversy.

A successful tactic: The acquired distinctiveness of cues

One of the most influential tactics in resolving the continuity-noncontinuity controversy was D. H. Lawrence's method of demon-

strating that rats learn to attend to particular types of cues. He called this the "acquired distinctiveness of cues." Essentially, he trained rats to respond to a particular type of cue, then assigned them a new task in which the previously learned *responses* (approach and avoidance tendencies) were irrelevant but knowledge of the relevant *type of cue* was important. In order to do this, he trained rats on a simultaneous discrimination, then shifted them to a successive discrimination.

A simultaneous discrimination is one in which the correct and incorrect cues are both present on any given trial. The task of the organism is to approach one of the cues, designated as the correct one. With a successive discrimination, only one cue is present on a given trial. For example, if black is present on a trial, this may mean "go to the right," and when white is present, that may mean "go to the left." Notice that the animal does not learn to approach a given brightness (say, black) or to withdraw from the other brightness (say, white) when doing a successive discrimination.

Lawrence (1949) trained groups of rats, each with a different type of cue, on a simultaneous discrimination. One group discriminated black from white, a second group discriminated rough from smooth alleyways, and a third group discriminated wide from narrow alleyways. Having learned the simultaneous discrimination, they were presented with a successive discrimination. As a concrete example, take animals that have learned a simultaneous discrimination of black from white. Now the same animals encounter a successive discrimination with black and white as the relevant cues. From the point of view of noncontinuity theory, they will do well on the successive discrimination because they have already been trained to attend to the relevant dimension. But continuity theory predicts no advantage for them. This is because the approach tendencies learned in the earlier task are of no value. On any given trial either black is on both sides or white is on both sides. The tug to right or left produced by approach tendencies should be equal.

The two theories also make different predictions about the effects of earlier training on dimensions irrelevant to the successive discrimination. If an animal has been trained to discriminate rough from smooth, noncontinuity theorists would say that it has learned to attend to texture and ignore such cues as brightness. This should impair learning on a successive brightness discrimination. Once again, continuity theory predicts no effect because approach and avoidance tendencies provide equal tugs and pushes for the right and left.

Lawrence found that prior black-white training aided in the later successive discrimination with black and white as the relevant cues. This seems clearly to support noncontinuity theory. On the other hand, the irrelevant cues did not seem to interfere. But null acceptances are

not as weighty as positive findings, so noncontinuity theory gained ascendance after this study.

Other relevant evidence

No single study could really convince us that noncontinuity theory is correct. We rely on the cumulative evidence from various studies. Evidence mounted over the years in favor of the noncontinuity point of view. For example, Reid (1953) showed that animals trained to a learning criterion, then given hundreds more training trials, actually learned a reversal of the original discrimination faster than animals only trained to criterion. This is called the *overlearning reversal effect.* A continuity theory had a hard time explaining it. Overlearning (the extra trials after criterion) should build up the original tendencies to approach and avoid. The greater the overlearning, the greater should be the original tendencies, and the more trials it should take to break them down and learn the reversal. Noncontinuity theorists can account for the finding by pointing out that the organisms have learned to attend to the correct type of cue. Reversing their conditioned approach and avoidance tendencies is a relatively small matter. This and many other findings have led to a wide acceptance of a noncontinuity point of view. In fact, we now find models of discrimination learning that have the quantitative precision of Spence's (1936) model (and more!), but that assume not just two but many stages of learning (see, for example, Lovejoy, 1968).

It is not hard to explain why certain studies supported the continuity view. Whether or not it will be possible to detect a phase of learning to attend to the correct cue dimension will depend on whether that dimension is one the animal attends to without any training. For example, rats tend to pay attention to spatial cues. It is hard for them to learn that visual details are important. With a spatial cue defined as the relevant one, little or no time will be spent on the stage of discovering the correct type of cue. Thus, the behavior will seem consistent with a continuity view. Small details of the apparatus can cause major changes in the readiness with which organisms notice a given cue (see, for example, Meyer, Treichler, & Meyer, 1965).

MORE RECENT APPROACHES TO THE STUDY OF LEARNING

I have already mentioned some methodological changes that have occured since the heyday of the continuity-noncontinuity controversy. Many investigators continue to pursue the subtle questions of conditioning and learning within a framework and with methods more

Key Ideas Box 7.3: Some tactics used to resolve the continuity-noncontinuity controversy

Some major tactics used to resolve the continuity-noncontinuity controversy were:

1. The tactic of *presolution cue substitution*. While subjects are still responding at chance levels, the "correct" and "incorrect" labeling of cues is reversed. Correct becomes incorrect and incorrect becomes correct. Continuity theory predicts impaired learning as a result of reversal. Noncontinuity theory says that the organisms are not yet noticing the correct type of cue, so no impairment should occur. Different studies gave differing results, so the tactic failed to resolve the issue.

2. The tactic of *postsolution substitution of cues*. After a discrimination has been learned, noncontinuity theory says the organism's attention is riveted to the correct cue. An extra cue also correlated with reward, if slipped in at this point, will not be noticed. That extra cue will not influence the behavior later if the original cue is removed. For example, animals were trained to respond to the larger of two circles. Later a large triangle replaced the large circle. Noncontinuity would say that nothing would be learned about shape, since animals would pay attention to size. Later, if tested with triangle and circle of equal size, responses should be at chance levels. Continuity theory would say that the shape would gain control of the behavior, and correct responding would occur during the test phase. Two studies obtained opposite results, so, again, the issue was not resolved.

3. The tactic of *acquired distinctiveness of cues*. Rats are trained to respond to a particular type of cue, then assigned a new task where the previously learned *responses* (approach and avoidance tendencies) are irrelevant, but where knowledge of the appropriate type of cue is important. For example, rats trained to approach a black cue and avoid a white one are shifted to a task requiring them to go to the right when black is present on both sides and go to the left when white is present on both sides. The tugs and pushes of previously learned approach and avoidance tendencies are equal on any given trial. But knowing that brightness is the relevant dimension is helpful. This finding yielded results supporting a noncontinuity position.

4. *Other methods*. Several other methods, none of them sufficient alone, pushed investigators toward the noncontinuity view.

sophisticated than, but much like, those of the workers just discussed. Many people feel a certain frustration at such work, seeing it as "seeking knowledge of more and more about less and less." However, we must seriously consider the possibility that psychology, like other sciences, will have to become very subtle and deal with minutiae before it will be a true science of behavior. I suspect that any student of learning would have to confront the basic issue of continuity versus noncontinuity in learning.

Many students of learning and memory have shifted their attention away from such issues, however. In fact, there has been, in recent years, increasing enthusiasm for a different approach to these problems. The approach is called *human information processing.* It shifts emphasis from questions of the development of conditioned stimulus-response bonds over to questions of the manner in which humans process information. The widespread use of computers has provided a major impetus to this new approach. Parallels between the ways in which computers and humans process information are often striking.

The shift toward an emphasis on information processing is not merely a tactical shift. It is a major strategic change. Indeeed, there is a bit of a struggle going on between proponents of the traditional approaches and those interested in information processing. The strategic importance of this new point of view makes it essential that we go into some examples of its strategy and tactics. I have chosen to include work on information processing in both learning and memory.[2] In fact, some of the work borders on the areas of perception and attention. The traditional subdivisions are not as useful within the framework of information processing as in the older framework. But enough of these preliminaries. Let's look into human information processing.

Human information processing, learning, and memory

In Nathaniel Hawthorne's classic novel, *The Scarlet Letter,* Hester Prynne had a child out of wedlock. She lived in Puritan New England and was publicly punished for her sin. Part of her punishment was to endure the disgrace of wearing a scarlet "A" over her breast as public acknowledgement of her adultery.

Arthur Dimmesdale, who had fathered the child, seemed, on the surface of it, to have escaped punishment. But he carried an unbearable burden of conscience, for he was the minister of the very Church

[2]Learning and memory are quite different. When we learn we gain information, and change our response probabilities. With memory, we retain what we already acquired. Certain variables may influence learning and memory in quite different ways. For example, damage to the brain might result in faulty memory with ability to learn virtually intact.

that condemned her. The weight of his guilt soon caused a deterioration in him, which led to his resolve to leave the Church behind and run away with Hester and their daughter, Pearl.

Having so decided, he found himself tortured by diabolical urges too long suppressed. At one point when he was to console a pious and elderly woman of the Church, he leaned to whisper in her ear, but " . . . could recall no text of Scripture, nor aught else, except a brief, pithy, and, as it then appeared to him, unanswerable argument against the immortality of the human soul." Hawthorne states that the instillment of such an idea in her mind " . . . would probably have caused this ancient sister to drop down dead at once, as by the effect of an intensely poisonous infusion." But instead, as the minister, walking away, looked back, " . . . he beheld an expression of divine gratitude and ecstasy that seemed like the shine of the celestial city on her face, so wrinkled and ashy pale."

This is a nice example of exceeding the "information-handling capacities" of a human brain. The woman just couldn't grasp the unlikely message she got, and interpreted it as something more familiar. The information contained in a signal depends on its probability. If every morning for ten years George Boring has hailed Sam Gray with "Good morning, how are you?" and Sam has replied "Fine, how are you?" little information has been conveyed. Anything highly predictable is poorly informative. Unlikely signals are rich in information. So if Arthur Dimmesdale had quoted scripture, the information content would have been low. When he said something grossly improbable for a minister, the information contained in the signal was very high. Information content is inversely related to the probability of a signal.

George Miller (1956) showed, in a classic paper entitled "The magical number seven, plus or minus two," that the information-handling capacity of the human brain is quite low. Although an immensity of information is coming into the human brain from the thousands upon thousands of sensory input units, the brain can only process about 2.5 "bits" of information. A "bit" is a unit of information, independent of the particular dimension containing the information. So, for example, you can measure auditory, visual, or tactile information in the same unit, the bit.

What happens when too much information deluges us? We say that our "channel capacity" has been exceeded. The signal fails to get through. The information is lost. And, incidentally, we usually feel uncomfortable when our channel capacity is exceeded. When Professor Nitrick gives a long-winded, complicated lecture on the orbitals of electrons, we not only fail to understand what he said; we usually feel jittery, irritable, and bored. Stanley Milgram has suggested that peculiarities in the behavior of people in large cities might largely be due to

00 = A
01 = B
10 = C
11 = D

First memorize the new code very thoroughly. Then repeat the original experiment, determining how long a string of digits you can recall when you recode each pair. This means that you will listen to the numbers as they are read to you and give the first pair a letter symbol, the next pair a letter symbol, and so on. For example, the string given above becomes: B, C, C, A, C, D, C. Now all you have to remember is the list of letters, which you can then decode in order to repeat back the original numbers. The result is a great increase in your memory span.

Further improvements can be made by recoding strings of three digits (000, 001, and so on), or even longer strings. More time has to be taken in memorizing the code, but once done, the feats of memory you can perform are prodigious indeed! Smith (cited in Norman, 1969) showed that with a 4:1 or 5:1 encoding system he could repeat back as many as 40 digits without error.

STAGES OF INFORMATION PROCESSING

Notice that as you perform this little exercise in memory, you do it in definite stages. You must learn the basic code, turn your attention to the digits, recognize the digits, recode them, rehearse the letters internally, decode the letters into their appropriate digits, and finally read out the list. You are going through stages of information processing. You have to go through many, if not all, of these stages when using natural memory. Some of the stages may take place at the same time. Others can only occur one after the other. For example, you may be able to focus your attention at the same time as you recognize the digits. But you must recognize the digits before you can recode and rehearse them. If two stages can occur simultaneously, we say that *parallel processing* is taking place. If they must occur one after the other, it is called *sequential processing.*

The idea of stages is a key one for those who study human information processing. So it is important to know how to identify a stage. What makes us decide whether there are two stages or one in a given processing task? Simply this: a stage is identified by characteristic functional properties. For example, we might find that certain aspects of processing information take more time than others. Or we might find that there are great variations over time in the channel capacity for certain information. Sometimes we find that the dimensions of encoding vary over time. For example, we sometimes code verbal information according to how words sound. At other times we

encode it according to what they mean. If we examine how a person handles information and find that these functional properties vary over time or that they vary within a given temporal unit, depending on which aspect of the information is being processed, we conclude that there are different stages of information processing.

The study of human information processing focuses on the delineation of stages and on identifying their functional properties. It provides a general framework that can actually be extended to all of psychology (see, for example, Massaro, 1975). As a conceptual framework for psychology, it is having a very large influence. Perception, attention, thinking, and memory have been placed within it (Haber, 1969; Norman, 1969).

A classic experimental attempt to break down information processing into stages was that of Donders (1868–1869, translated 1969). When measurements are made on a phenomenon, science begins to encompass that phenomenon. Donders wondered whether it might not be possible to measure the time required to shape a concept or express one's will. In effect, he tried to differentiate stages of processing by measuring the amount of time taken in each stage. By devising tasks of increasing complexity he hoped to determine the time taken for the new components of more complex acts by a *method of subtraction.* Later, this became the basis of *mental chronometry,* which dominated the work of Wilhelm Wundt's laboratory for some time. Let's examine mental chronometry and the method of subtraction.

Reaction time

The basis of the method of subtraction is the measurement of *reaction time.* This is the amount of time taken for an observer to respond to a given stimulus or set of stimuli. Reaction time first became a subject of scientific interest when a Mr. Kinnebrook was fired from the Greenwich Observatory in 1796 by its director, Maskelyne. Maskelyne had noticed that Kinnebrook's recorded times for the locations of stars were about 0.5 sec later than his own. This was an important error, because all timepieces in the country were standardized on Greenwich time. Maskelyne warned Kinnebrook to improve his performance, but to no avail. Kinnebrook was dismissed on the suspicion that he had fallen "into some irregular and confused method of his own" instead of using the established method of recording time.

The incident was recorded in *Astronomical Observations at Greenwich,* and came to the attention of the astronomer Bessel, who was interested in error of measurement. Bessel doubted that Kinnebrook would persist in a method that caused him so much trouble, and began investigating the source of error. Essentially Bessel found that a certain amount of time is required for reacting to a stimulus, and that the amount of time is considerable. Furthermore, it varies, especially

across individuals. To get agreement on temporal judgments among observers it was necessary to correct for these variations in reaction time. The correction came to be called the *personal equation.*

The personal equation was obviously a psychological one, and psychologists soon became deeply involved in the study of reaction time. Donders, who was actually a physiologist interested in vision, used reaction time as the basis for his mental chronometry. With a simple reaction-time task the observer must make a response such as pressing a button as soon as possible after a stimulus has been triggered. The stimulus may be a light, a tone, a touch, or what have you.

The reaction times for stimuli from these different sensory systems will be different from each other. But, staying within one sense, Donders complicated the task by adding other mental processes. To add *choice* to the task, he set up several stimuli, each with its own appropriate response. For example, he would stimulate either the right foot or the left foot. The subject was to respond with the right hand if the right foot was stimulated and with the left hand if the left foot was stimulated. By subtracting the reaction time involving choice from the simple reaction time, he believed he could measure the time taken for the mind to make a choice.

Later, Donders decided that there were really two things being added with his more complex task. These were choice and discrimination. So he devised a method for extracting out the time taken for discrimination. He used a number of stimuli, A, B, C, and D. But the subject was only to respond to stimulus A with response A. By subtracting, Donders seemed able to get times for choice, discrimination, and reaction.

EVALUATION OF DONDERS'S METHOD. In the end, Donders's method failed. For one thing, it is probably impossible to eliminate choice from a task. Even when there is only one response, we choose whether to respond or not respond. Furthermore, subtraction cannot be relied upon. When one stage is altered, all the other stages may, and probably do, undergo change. The elements interact. We cannot say that the total time is equal to the sum of the component times, any more than we can say that the volume of a jar full of marbles plus the volume of a jar full of sand may be added to yield the volume of a jar in which the sand and marbles are mixed.

Contemporary analysis of stages of processing

Stages are identified by properties other than their duration. In fact, we are less interested in their duration than in such things as the transformations of information that take place. What properties of stimuli are taken into account during a given stage? Do these proper-

Key Ideas Box 7.4: Information processing

The entire process of receiving stimuli, attending selectively, storing in memory, deciding how to react, and checking on the adequacy of the reaction may be regarded as information processing. Such processing seems to occur in *stages,* each of which has its own functional properties. A given stage may take more or less time than another, may rely on different features of the stimulus than another, and so on. Students of human information processing want to identify the stages and their properties. Two stages may take place one after the other *(sequential processing)* or both at the same time *(parallel processing).*

Early work on information processing (even before it went by that name) used *reaction time* to identify the stages. Reaction time is the duration of time between presentation of a stimulus and the response to it. By varying the nature of the task so that it did or did not involve choice, discrimination, and the like, experimenters hoped to time such processes. Their method was called the *method of subtraction* because it entailed subtraction of simple from more complex reaction times. The method of subtraction failed because it was impossible to eliminate certain mental operations and because a change in one part of a task tended to modify all parts of it. Fortunately, stages can be identified by properties other than their duration, and these are the focus of contemporary work.

ties differ from stage to stage? What stratagems are used in scanning information in a stage? How is information lost in a stage? Questions like these, which are independent of Donders's method, interest us now.

HUMAN MEMORY

The study of memory has played a key role in history. In ancient times, when printing and writing were not readily available, great demands were made on memory. It is said that Themistocles knew the names of 24,000 citizens of Athens. Such feats are commonly reported on the part of the ancients. There are many fascinating links between the ancient arts of memory and such things as astrology, Sufism, and the numerology of the Cabala (see Yates, 1966). But the experimental psychology of memory seems to have begun with Herman von Ebbinghaus.

Ebbinghaus, using himself alone as subject, devised a rigorous

Herman von Ebbinghaus

method of measuring memory. He also determined many functions relating learning and memory to such things as the amount of material to be learned, the time elapsed since learning, and the number of repetitions. For example, he gave us the typical "curve of forgetting," which declines rapidly at first and then makes further declines very slowly for a long time.

The results of Ebbinghaus's work were published in a brilliant little book called *Memory, a contribution to experimental psychology,* which is available in English (Ebbinghaus, 1885). It was a great contribution to the general trend toward mental measurement. He, like so many others of his time, measured that which was believed to be unmeasurable, and thereby strengthened belief in the possibility of a science of psychology.

The work of Ebbinghaus was obviously a scientific breakthrough of great magnitude. He started a tradition of research on memory that is clearly recognizable in modern-day psychologists who are interested in memory: A great deal of modern work could rightly be regarded as a series of footnotes to Ebbinghaus.

Any scientific discoverer is likely to make a number of methodological innovations, and Ebbinghaus was no exception to this rule. One major tool of his devising was the nonsense syllable. Ebbinghaus was steeped in the powerful analytic traditions of nineteenth-century science, and was naturally concerned to see that his methods included the highest possible degree of control over extraneous factors that

might influence memory. Furthermore, a scientist trained in his tradition could be expected to insist on the greatest possible simplification of experimental materials. Mature sciences had succeeded by finding the laws governing the behavior of very simple systems under highly controlled conditions. Many scientists even today regard such analysis as essential to a truly scientific procedure. The simplest situations must be studied first, and the more complex ones either studied later in the light of simpler laws or, preferably, interpreted by combining simple laws. To be sure, an alternative view emphasizes the need for direct study of complex systems as a whole (see, for example, von Bertalanffy, 1968; Weiss, 1967; or even Bartlett, 1932), but this is not the time to deal with that issue. No one can deny that the analytic tradition of Ebbinghaus was and is a fruitful one.

How did Ebbinghaus simplify his materials while continuing to study a "higher mental process"? He used nonsense syllables as the material to be learned. His nonsense syllables were meaningless combinations of three letters, each combination consisting of two consonants separated by a vowel or diphthong. For example, *bap* and *zup* might be a typical pair among such nonsense materials. It is, of course, possible to make nonsense syllables according to other rules. For example, fewer or more letters could be used; consonants or vowels alone could be employed. Three-letter syllables like those of Ebbinghaus are still commonly used, and they are called *trigrams*.

What did Ebbinghaus gain by using nonsense syllables? For one thing, he simplified his learning task so that it had greater promise of being amenable to analysis. Further, he presumably reduced variability by reducing variations in the familiarity and meaningfulness of his materials. If he had used poems, sections from plays, or other common material, major variations in the degree of previous acquaintance with the material would have been present, and the meaningfulness of the material would also be subject to uncontrolled variation. Not only did Ebbinghaus hope to reduce such variations in general, but also, more specifically, he saw the necessity of having access to equivalent materials for repeated tests. This was necessary because Ebbinghaus used a within-subjects design (in fact, Ebbinghaus himself was the only subject). Ebbinghaus did certain tests that encouraged him to believe that the results with nonsense syllables could be generalized to meaningful material.

Free-recall, paired-associate, and serial learning

We commonly use one of three types of task to study verbal learning and memory. A *free-recall* task requires that the learner memorize a list and recall it in no particular order. *Paired-associate* tasks require linking of a stimulus syllable with a response syllable. For example, *yub* might be the stimulus word and *gib* the response word. The task

would be to respond with *gib* when presented with *yub*. Normally, the task entails memorizing a large number of these paired associates. With *serial* learning tasks, the learner must memorize a list in a specified order.

The association value of nonsense syllables

One way to look at Ebbinghaus's nonsense syllable technique is to think of it as a method of getting rid of associations. Ordinary words have associations to varying degrees. Since these will vary in an uncontrolled way, it can be of value to eliminate them. This is what the nonsense syllable is designed to do.

But does the nonsense syllable really get rid of associations? No. Nonsense syllables do in fact call to mind other verbal contents. It simply does not appear possible to eliminate associations entirely. This is not an atypical problem in science (see Chapter 4). One often has difficulty lowering the value of a given variable to zero. There is a tendency for such attempts to result in values that are simply unknown.

A good way to cope with problems of this sort is to *manipulate* the variable in question. This is exactly what students of verbal learning have done. A number of techniques have been devised for determining the association values of nonsense syllables. For example, a nonsense item might be presented to a subject who has been asked to list as many things as the item brings to mind. If a fixed period is allowed for making such associations, then some measure, such as the proportion of subjects able to think of any association at all, can be used to rate the association value of the item. For example, if no one in a group of subjects can think of an association with the item, it may be said to have a zero percent association value; if half of the subjects can think of an association, it may be said to have a 50 percent association value, and so forth. Other, similar techniques have been used for deriving association values of syllables, and these are available in standard tables (see, for example, Underwood & Schulz, 1960).

Having available lists of nonsense syllables with known association values makes it possible, among other things, to come closer to the goal of Ebbinghaus by finding lists that have been empirically demonstrated to be very low in association value. In addition to this, the association-value variable (commonly referred to as "meaningfulness") can be manipulated in experimental contexts in order to find functional relationships based on it. This has been done, for example, by Underwood and Schulz (1960).

There are a number of known relationships between such "meaningfulness" and other variables. For example, there is a relationship between association value and the ease with which a syllable can be pronounced (pronunciability). Probably related to this is the obser-

vation that high association value improves rates of learning nonsense-syllable lists if the high association-value syllable is the response item of a paired associate, but not if it is the stimulus item.[3]

This paradoxical finding has been taken to mean that paired-associate learning consists of two phases: the response-learning phase, in which the learner acquires the ability to make the response; and an association phase, in which the subject links the stimulus and the response together. Since high association value is correlated with high pronunciability, response learning is facilitated by having response items of high association value, but not by having stimulus items of high association value. Savings in the response-learning phase of paired-associate learning are thus found when the response word has high "meaningfulness," but not when the stimulus word is similarly high in association value (Underwood & Schulz, 1960).

Evocation of imagery by nonsense syllables

In recent times experimental psychologists have come to recognize the importance to verbal learning of a discredited mentalistic concept. The concept is that of imagery (Paivio, 1971). In fact, we do not have to abandon our dedication to measurement in order to include imagery in our science. We have only to specify measurement operations for imagery. For example, suppose we ask subjects to rate the extent to which a syllable tends to evoke an image. The resulting number is a legitimate operational definition of "imagery."

The ease of memorizing paired associates is related to their tendency to evoke images. Pairs of syllables that readily evoke images are also readily memorized. But imagery, like pronunciability, is positively correlated with association value. What happens if we control for imagery while holding association value constant, or vice versa? Paivio and his colleagues did this in a series of experiments on free-recall, serial, and paired-associate learning (see Paivio, 1971). Though all the returns are not yet in, it seems that association value has relatively little effect in the absence of imagery, whereas imagery is influential when association value is held constant.

CONTRIVED VERSUS SPONTANEOUS IMAGERY

When Ebbinghaus first set out to measure memory, he stirred excitement about the possibility of developing measures of the power of minds. One contemporary said, "May we hope to see the day when school registers will record that such and such lad possesses 36 British Association units of memory-power or when we shall be able to calculate how long a mind of 17 'macaulays' will take to learn Book II

[3]There are, of course, limiting conditions on this finding (Paivio, 1971).

of *Paradise Lost*" (Hilgard, 1964). Today we have IQ tests, but the results do not predict success in learning to the extent anticipated by Ebbinghaus's contemporaries. There are many reasons for this, but a major one is that strategies of information processing tend to outweigh "basic brainpower" in determining learning.

When Ebbinghaus measured his ability to learn lists of nonsense syllables, he found that he could memorize about 7 syllables in one trial. Ebbinghaus was brilliant, but he differed little from the average person in this basic ability. A much less brilliant person who had studied the ancient arts of memory could do far better. The exercise described earlier, in which lists of binary digits are re-encoded in order to extend the learner's span of memory, illustrates this. Chunking of information is of great importance. Imagery is also used by "mnemonists," as specialists in the art of memory are called. But instead of relying on spontaneous images, they arrange special procedures to encourage the evocation of images. The following exercise provides a small example of the method.

THE USE OF CONTRIVED IMAGERY: AN EXERCISE AND AN ILLUSTRATION This can be done in the style of Ebbinghaus, as a within-subjects design with yourself as subject. If you have more subjects, use a within-subjects design with counterbalanced order of presentation of the independent variable (instructions to use or not use associative imagery).

First you must memorize the following list:

one is a bun
two is a shoe
three is a tree
four is a door
five is a hive
six are sticks
seven is heaven
eight is a gate
nine is wine
ten is a hen

Be sure the list is firmly in memory before beginning. The experiment requires that you learn in correct serial order lists of ten words selected haphazardly (say, from the dictionary). With half of the lists, try to create an image associating the words in the list above with the new words. For example, if the first word in your new list is "house," you might picture a gigantic hamburger bun with a house inside it, or a hamburger stand built to look like a bun, or a suburban BUNgalow. If the next word is "bird," you might picture Donald Duck putting a shoe on, or the winged shoes of myth and fairy tale, or even a scarecrow

"shooing" birds. On the other half of the trials, refrain from making such associations. If, after having learned the system, you have trouble refraining from using the images, try a completely randomized design in which some subjects haven't learned the list. There really should be a control in which they learn some other, unrelated list. Do you find the artificial associations effective?

Eidetic imagery

While discussing the relationship of imagery to memory, a word or two should be said about the phenomenon of *eidetic imagery*, popularly regarded as *photographic memory.* Eidetic images seem more like real perceptions than do the usual memory images. The eidetic images appear to be outside the person, as though they were being seen. The topic of eidetic imagery interested early psychologists (see review by Klüver, 1932), but fell out of favor as psychologists came to spurn mentalism. More recently, Haber and Haber (1964) revived interest in it.

Such imagery is said to occur more often in children than in adults, more often in retarded children than in normal children (Siipola & Hayden, 1965), and more often in nonliterate people than in literate people (Doob, 1964). Though we usually expect people with such abilities to image to have unusually accurate memory, this is not always the case (Oswald, 1960). Some eidetic imagers see vivid but, apparently, inaccurate images. Yet Haber and Haber (1964) noted a degree of superiority in memory for visual details among eidetic children. In some cases eidetic imagery and stupendous memory go hand in hand. Stromeyer and Psotka (1970) reported results of observations using dot patterns composed of many thousands of dots (Julesz patterns). The patterns were designed so that normal subjects can, if given one pattern in one eye and a second in the other eye, fuse the two into a third image. The subject of Stromeyer and Psotka (1970) could fuse two patterns presented sequentially, even with days intervening between presentations.

ASSOCIATIVE BONDING

A major interpretation of verbal learning centers on the notion of associative bonding. Responses become attached to stimuli by strengthening of the bonds between them. The similarity of this process to that of respondent conditioning is clear; the idea of association is an ancient one. Interpretation by appeal to associative bonds applies most directly to learning of paired associates, where there is a stimulus "word" and a response "word." The stimuli are less readily identifiable with, say, free recall. But the idea of learning by

Key Ideas Box 7.5: Tactics in the study of verbal memory

Ebbinghaus first measured human memory. He used repeated measures on only one subject, himself. He devised *nonsense syllables* in order to have uniform material from one time to another and to minimize the variability due to mental associations with verbal material.

Contemporary psychologists have refined the tactics devised by Ebbinghaus. They have found that there are substantial variations even in the associations to nonsense syllables, and have studied the effects of these variations systematically. For example, learning is commonly faster if verbal responses have high *association value.* Association value is correlated with other properties of verbal material. Materials high in associations are easier to *pronounce* and evoke more *images.* There is evidence that the major effects of association value are due to this evocation of imagery. Study of the influence of imagery on memory is widespread today. This includes the study of spontaneous imagery as well as the use of contrived imagery to improve memory. *Eidetic imagery,* which is vivid and appears to be outside one's head, has also been studied. It is not necessarily related to exceptionally accurate memory.

Three major types of task are used in the study of verbal learning and memory. They are *free recall,* in which memorized lists are recalled in any order; *serial learning,* in which memorized lists must be recalled in stipulated order; and *paired-associate* learning, in which a stimulus and a response "word" must be linked.

formation of stimulus-response bonding can be used as an interpretation of virtually all the phenomena of verbal learning and memory.

Interference is a concept that is especially important in applying the stimulus-response paradigm to human learning. It enables us to account for such things as transfer of training and forgetting. *Transfer of training* refers to the influence of prior learning on the learning of a new task. Transfer may be either positive or negative. When I or any other trained psychologist read technical books in psychology, positive transfer is exemplified. I can learn the material very quickly. Often it is easier for me to grasp information in a technical book than in a novel.

A friend of mine once displayed negative transfer. He was a weightlifter and wanted to learn to shoot pool. Weightlifting taught him to put a lot of muscle into what he did. So he fell upon the strategy

of shooting pool with all his might. A lot of balls go in that way, but it is not the way to become a good pool player.

Many psychologists believe that forgetting is entirely due to interference. They believe that memory traces do not simply decay. When we forget, it is because other stimulus-response bonds interfere with what we learned. Two basic types of interference influence forgetting.

Proactive interference occurs when prior learning interferes with later learning. I said earlier that I could learn new technical material in psychology very quickly, due to positive transfer. But I also forget it very quickly (even things I write myself!) This aspect of the absent-minded professor syndrome is due to interference by the many things I learned earlier about psychology.

An experiment on verbal learning designed to test for proactive interference might look like this:

Experimental Group	Learn List A	Learn List B	Test for Retention of List B
Control Group	Rest	Learn List B	Test for Retention of List B

In the experimental group, proactive interference is present. In the control case, no list is given to create proactive interference.

A second major kind of interference is *retroactive interference.* Having learned some material, if I go on to study related things the newest learning will tend to interfere with previous learning. This is one reason students are often advised to go right to bed after studying for an exam to be taken the next morning. If they read magazines, talk, and so on after studying, it speeds the forgetting process.

A typical experiment on retroactive interference looks like this:

Experimental Group	Learn List A	Learn List B	Test for Retention of List A
Control Group	Learn List A	Rest	Test for Retention of List A

In the experimental group, retroactive interference is present. In the control case, no list is given to create proactive interference.

Not all psychologists agree that all forgetting is due to interference. In fact, there is tension between older interpretations that place exclusive emphasis on interference and the newer interpretations in terms of stages of information processing. For example, one of the early models of information processing, that of Donald Broadbent, represented loss of short-term memories as a process of *decay* rather than interference (Broadbent, 1957).

SENSORY MEMORY AND THE METHOD OF PARTIAL REPORTS

There is an old literature in psychology on what is called the *span of apprehension.* When you briefly perceive an array of stimuli, how

Key Ideas Box 7.6: Associative bonding and interference

A major way to interpret learning and forgetting is to suppose that stimuli are becoming *associatively bonded* to responses. Old bonds may interfere with the formation of new ones. This is called *negative transfer* of training. They may also interfere with memory traces. This is called *proactive interference.* Old bonds may also facilitate new learning. This is *positive transfer* of training. Sometimes the memory for a task already learned may suffer interference from a related task learned later. This is called *retroactive interference.* Some psychologists feel that all forgetting is due to interference. Though other psychologists argue that there is also a process of *decay* of memory, it has been difficult to find an experimental example of a memory loss better explained by decay than interference.

much of the information can you take in at a glance? Surprisingly little. For example, take the following array of stimuli:

KNLF
TSNZ
RKLT

If it were visually presented for about 50 msec (milliseconds, thousandths of a second), you would likely be able to repeat back only four or five of the letters.

But how much of the information did you lose while in the act of telling what you saw? We know that a good deal of information can be lost in very brief time periods, so taking the measurements in the way just described probably yields an underestimate of the actual span of apprehension. Sperling (1960) devised an ingenious method of getting a more accurate measurement. This is called the *method of partial reports.* Immediately after the stimulus array was turned off, he signaled with a prearranged tone that indicated which of the rows was to be remembered. Subjects could then focus their attention on the particular information to be reported back, but their only source of information was the memory image.

The result was that subjects could report back almost 100 percent of the letters. Sperling interposed various delay times prior to the signaling tone, and found that accuracy dropped sharply with time. With a one-second delay accuracy was down to the ordinary levels of earlier research. This kind of memory trace, called the *iconic* memory, appears

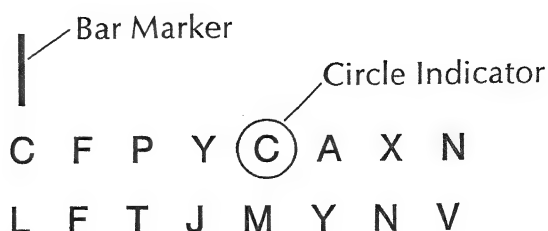



FIGURE 7.3 Illustration of the type of display used by Averbach and Coriell (1961) to measure iconic memory. Rows of letters were presented then a bar marker or circle indicated which letter was to be recalled. Backward masking tended to occur when the circle was used. See text.

to be a rapidly decaying persistence of the percept. Subjects report that it seems to still be there as a stimulus. It is not experienced as a memory image, though it really is one.

A good deal of further research has been done on iconic memory. Averbach and Coriell (1961) used a similar method that leads to some interesting complexities. They presented rows of letters and signaled which individual letter was to be recalled. Instead of a tone, they used visual signals, either a pointer or a circle surrounding the appropriate letter. Figure 7.3 illustrates their procedure. By and large, their results were much like those of Sperling. However, they found that *backward masking* occurred, especially when the letter was surrounded by a circle. With backward masking, the perception of the letter is obliterated by the subsequent circle. A good deal of earlier research had been done on backward masking by various investigators, but the work of Averback and Coriell seemed to provoke a great deal more. This work is discussed in Neisser (1967).

Incidentally, the duration of the iconic memory is not always one second. Its duration depends on such variables as the duration of the stimulus, its intensity, and the nature of the backdrop. For example, if subjects are in the dark during the retention interval, the icon may last as many as five seconds.

SHORT- AND LONG-TERM MEMORY

I know you've experienced it time after time. You go to a party. You walk in the door and right away someone says, "I'd like you to meet Sam Brady." You put out your hand and smile, "Hi Sam." You know you're going to be meeting a lot of people tonight, so you look carefully at Sam and say to yourself, "Sam Brady, Sam Brady," trying to fix the name in memory. Then Sam notices somebody, calls you over and says, "This is Mary Gordon." Off you go again, "Mary Gordon, Mary Gordon." You talk for a few minutes, then along comes your old friend

Key Ideas Box 7.7: Sensory memory and the method of partial reports

For a brief period after presentation of a visual stimulus it persists in the form of an image or *icon*. Virtually all of the information in the stimulus is available during that time. This will not show up if subjects are merely asked to report on what they saw. Forgetting occurs so quickly that part of the information is lost while reporting. Sperling devised a *method of partial reports* that reveals the extent of the information storage. With this method a relatively complex array of stimuli is presented, but immediately after it is shut off a signal indicates which part of the array is to be recalled. Recall is then excellent.

Morty Hilbert. "Hi ya, Morty," you say with a smile. Then a clutched feeling comes over you as you realize you can't introduce Morty to the two new acquaintances because you have forgotten their names.

Undoubtedly you got past the stage of sensory memory. But why couldn't you remember the few words involved in the names of a couple of people? A very important study by Peterson and Peterson (1959) shed some light on the problem. They gave their subjects a single trigram to remember and tested them over intervals of seconds, yet they found poor retention. They were able to get this bizarre result by preventing rehearsal during the interval of retention. Following each trigram, they presented the subject with a three-digit number. The subject was required to begin immediately counting backwards by odd numbers, starting with the presented three-digit number. Subjects were thus prevented from rehearsing, and the information was generally lost within 18 sec from the presentation. This parallels what happens when you are introduced and immediately distracted from rehearsing by having to carry on a conversation or meet another person.

In general, we seem to start with a percept, keep it briefly in sensory memory, then begin rehearsing it. Even if the presented stimulus is visual we tend to convert it into something acoustic. We repeat it over and over to ourselves. Given enough of this rehearsal, it goes into long-term storage.

The distinction between short- and long-term memory has been made on many different grounds. None of them is entirely convincing (Wickelgren, 1973), but the converging evidence is quite strong. Besides the sort of evidence provided by Peterson and Peterson (1959), there is evidence stemming from the method of free recall. If a subject is asked to memorize items from a list and repeat them back in no

particular order, there is a systematic pattern in the recall. There tends to be a primacy effect and a recency effect. That is, the earliest and the latest items in the list tend to be remembered best.

The recency effect is of relevance here. It appears to be due to a separate short-term memory system. As we shift from short- to long-term memory, there tends to be a shift from sensory to semantic[4] features of the memorized material. Consistent with this interpretation of the recency effect is the finding that the recency effect occurs only with immediate free recall, not with later free recall. Retention of items late in the list is actually worse for the later items if the test of recall is given a substantial time after learning the list (Craik, 1970; Rundus et al., 1970).

Many students of information processing have supposed that short-term memory was subject to decay, whereas forgetting of long-term stores came about by interference. However, Keppel and Underwood (1962) demonstrated the importance of interference in experiments of the type devised by Peterson and Peterson (1959). They measured recall after either 3 or 18 sec, and they varied the number of items already encountered prior to testing on a given item. The number of preceding items varied from 0 to 5. When the item to be recalled was the first experienced by a subject, recall was almost perfect even after a sec. Recall declined as a function of proaction. Thus, the results of the Peterson and Peterson study were apparently heavily influenced by proactive interference.

THE LIMITS OF MEMORY WITHOUT REHEARSAL: AN EXERCISE AND AN ILLUSTRATION To get a feel for the Peterson and Peterson (1959) type of study, try the following. It will be convenient to have a partner. Consonant trigrams will be presented, followed immediately by a three-digit number. The experimenter should read the consonant trigram and the number in quick succession. The subject must repeat the number immediately and count backwards from that number, by threes if it is even and by fours if it is odd, until the experimenter says STOP. Be sure the subject counts as quickly as possible, and thus cannot rehearse. With the STOP signal, the subject stops and reports the trigram immediately. After a 5-sec rest period, repeat with another trigram-number pair. The subject should be tested a minimum of eight times at each of six retention intervals (3, 6, 9, 12, 15, and 18 sec). If results are still unstable at that point, do more. If possible, use more than one subject. A list of trigrams and numbers follows.

[4]"Semantic" means having to do with meaning. Thus a shift from the way a word sounds to what it means would be a shift from sensory to semantic features.

Key Ideas Box 7.8: Short- and long-term memory

Having received a stimulus, we commonly re-encode it according to its acoustic properties and rehearse it for a while until it goes into long-term storage. This is the stage of short-term memory. Peterson and Peterson (1959) did a classic study in which they showed that subjects would forget a single syllable composed of three consonants (trigram) within 18 sec if they were prevented from rehearsing in the interim. This suggested the importance of rehearsal for short-term storage and gave an indication of the duration of that stage.

Some investigators have questioned the need for separation of memory into short- and long-term stages. They argue that there is no strong evidence that the two stages function according to different principles.

BFH 415	HCT 794	GMB 923	KGD 526
PGZ 574	JDP 209	TQH 664	RPB 193
BJH 226	ZFT 306	JNB 515	MQX 609
QKG 512	MPB 717	HZT 678	ZSN 596
CJP 827	QSG 288	QSW 521	LXH 525
BGK 400	CMZ 619	ZBK 800	MBS 112
SHX 681	FNQ 528	KMX 335	JXC 227
DKM 200	SZP 575	CQK 758	NLR 284
TJN 532	DFX 700	SBF 237	HBK 136
CMF 644	NRZ 787	THL 523	DSZ 415
GJP 311	LSH 894	XFR 268	KPB 171
FDM 156	PXT 151	BPM 537	NCS 395

In the event that a sizeable number of subjects is available, vary the order of presentation of the intervals. With enough subjects at each interval, you can treat the data from each subject's first interval as though they were from a completely randomized design. Do the results resemble those of Keppel and Underwood (1962) for subjects receiving their first interval?

Study Questions

1. What does it mean to say that science progresses from the vague to the specific?
2. On what species did Thorndike do his experiments?

3. What is a puzzle box?
4. Describe the typical result obtained by Thorndike with the puzzle box.
5. Contrast Köhler's findings with those of Thorndike.
6. Describe Köhler's method of measuring insight.
7. How did Köhler account for the difference between his findings and those of Thorndike?
8. What general principle of psychological measurement is implied in Köhler's account of the difference between his results and those of Thorndike?
9. Why did Köhler's findings fail to show that insightful learning occurs without previous trial and error? What subsequent research is relevant?
10. Compare and contrast the continuity and the noncontinuity views of learning.
11. Explain Krechevsky's discovery of hypothesis testing in rats and show how this makes clear the importance of the measurement method in determining laws of behavior.
12. Explain and evaluate the tactic of presolution reversal of cues.
13. Explain and evaluate the tactic of postsolution substitution of cues.
14. Explain and evaluate the tactic of acquired distinctiveness of cues.
15. What do we mean by "information" and "information processing"?
16. Explain the following:
    a. chunking
    b. sequential processing
    c. parallel processing
    d. stages of information processing
    e. reaction time
17. What was the major significance of Ebbinghaus's book on memory?
18. Why did Ebbinghaus use nonsense syllables?
19. What is a trigram?
20. Distinguish between paired-associate, free-recall, and serial learning.
21. How are nonsense-syllable association values measured?
22. If a variable cannot be reduced to zero, what alternative does an experimenter have?
23. If you wished to vary nonsense-syllable association values, what would be the simplest, easiest way to get stimulus materials?
24. What is the relationship between "meaningfulness" and association value?

25. What influence does association value have on rates of learning nonsense-syllable lists?
26. What theoretical account of paired-associate learning is in consonance with findings concerning the relationship between paired-associate learning and nonsense-syllable association value?
27. How are the following related to verbal learning and memory?
    a. spontaneous imagery
    b. contrived imagery
    c. eidetic imagery
28. Explain the concepts of:
    a. associative bonding
    b. positive and negative transfer
    c. proactive and retroactive interference
29. How do we measure transfer and interference?
30. Explain the method of partial reports and the discoveries based on it.
31. Explain the Peterson and Peterson method of measuring short-term memory.

8 PERCEPTION Fundamentals

The problem of perception

We perceive not just because of external stimuli, but also because of properties of ourselves. Nor is our experience of the world a passive process. We interact with the physical stimulus to create an experience, and thus the physical stimulus is only one of many variables controlling our perception.

To clarify this point, let us look at the familiar comparison between eye and camera. The eye, like a camera, possesses a lens that is capable of being flattened or thickened, an accommodation equivalent to the camera's focusing by moving the lens. It is a reasonably light-tight chamber otherwise; at the rear, where the camera has its film, the eye has a light-sensitive layer, the retina, containing several latent pigments that differentially absorb certain wavelengths of light, much as a color film does.

In all these respects the eye operates like a camera. But consider how strikingly it differs from a camera. It has a computer attached to it, and the computer, in determining what the experience will be, does not simply take into account what is taking place at the retina, but rather it considers a wide variety of information coming from the muscles, the ears, and other parts of the body. It also considers past experience,

present motives, and even adds the characteristic impress of its own computing structure.

It is easy to convince oneself that visual experience is not the result of mere passive registration of the physical event by observing certain afterimages. If a fairly bright image is presented for a brief duration and then turned off suddenly, the image will persist for a short while after the light has actually been turned off. This persisting image is called the "positive afterimage." After the positive afterimage disappears, it will be followed by an image of the same form but of the complementary color. For example, a red stimulus will ultimately be followed by a green image of the same form, and this is the negative afterimage. The negative afterimage is primarily due to the heightened threshold for stimulation by the wavelengths of light that have just been impinging on the retina.

The important point, however, is not that positive and negative afterimages occur. It is that their size depends upon the distance from the observer of the backdrop against which they are viewed. If the observer looks at a nearby screen, the afterimage will seem small. If he looks some distance away, let us say at a wall, the image appears large. In fact, a principle expounded by Emmert in 1881 (Emmert's Law) says that the size of the afterimage is directly proportional to the square of the distance of the backdrop from the observer. Note that the retinal image does not change as the backdrop changes. The retinal image in this special circumstance must remain constant, but the visual experiences are not constant at all. Hence, vision is determined by something more than the nature of the physical energies impinging on the eye.

The assumption that we "see things as we do because they are as they are" (Koffka, 1935) is false, and we are therefore faced with the question, "Why *do* we see things as we do?" Or more generally, why do we see, feel, hear, or taste things as we do? These are the central problems of perception, and their importance can hardly be exaggerated, as I hope it will be possible to show. Are there principles or laws that determine which way, of all the possible ways in which we might perceive an event, we actually do perceive it? If there are such principles, do they have their origin in the environmental encounters of an organism or in its inherited structure, or in both? More specifically, what sorts of environmental events give rise to which kinds of perceptual events, and what details of the organism's physical structure are relevant to its perceptual processes? These are the kinds of questions we will consider in the area of perception.

Various organisms of the same species and even the same organism at different times may construe physical events in different ways. Look, for example, at Figure 8.1. This is what is known as a "gestalt-ambiguous" figure. When it is shown to a group of people, the group inevitably separates into a number of individuals who see an old hag, a

FIGURE 8.1 What do you see? This is a gestalt-ambiguous figure, which can be seen either as a young woman or an an old hag.

number who see an attractive young woman, and a few people who see neither and who substitute an unusual percept of their own. What is particularly striking about this figure is that many people cannot for some time, even after considerable effort and instruction, see the figure in the manner opposite to the one that was spontaneous for them. It is always interesting to observe two people sitting side by side, the same physical object in front of them, yet "seeing" two radically different things. Consider how differently two observers may perceive more complex events or events in which the discrepancy between observers is less readily detected.

There is a great future for the study of perception in the investigation of such complex processes. The work done on cues relevant to social perception and on other social determinants of perceptual processes is typical of the research needed on complex processes. In all our interactions with others we respond to a variety of cues, many of which we would be hard pressed to identify. But these cues can be identified, measured, and manipulated by the experimentalist so that their influence can be determined.

For example, such cues as size of the pupil of the eye (Hess & Polt, 1960; Stass & Willis, 1967), distance between speakers in conversation (Hall, 1966; Willis, 1966), and shape of face (Dixon, 1966) have very

important influences on our perception of others. Stass and Willis (1967) did a particularly interesting experiment on this topic. Male college students were asked to choose between two women with whom they were to "work closely" in an experiment. The women had been selected from a larger group of women on the basis of their being rated about equally attractive by a large number of observers of both sexes. The independent variable was artificial dilation of the pupil with atropine, and the experimental design included the precaution of counterbalancing the two women so that half of the time one had atropinized eyes and half of the time the other's eyes were dilated. It was found that dilation of the pupil very greatly enhanced a given woman's chance of being selected.

THE VALUE OF STUDYING PERCEPTION

By now you have probably come to realize that there are important and interesting problems to be treated under the heading of perception, but a few additional points need to be made. Since the time of the philosophical empiricists, it has been unnecessary to emphasize the important role of the senses in determining psychological events. The sense organs are key structures, and no psychology can be complete until sensory and perceptual processes are understood. Furthermore, the areas of sensation and perception overlap other areas of psychology to a degree far greater than the novice might expect. In fact, though we have made a case in the preceding paragraphs for the existence of a special class of problem that is characteristic of the study of perception, in practice it is often hard to decide whether a particular problem is concerned with perception, learning, motivation, emotion, or other areas. As with most terms in ordinary language, the term "perception" does not have a hard and fast meaning; yet it constitutes a useful heading for certain types of research problems. Though inexact, ordinary language does guide behavior—even the behavior of scientists.

It is no coincidence that the various areas of psychology, including perception, tend ultimately to meld into a kind of unity. We usually think only of what may be loosely termed the "informative" features of sensory inflow, but this hardly does justice to the facts. Perceptions also vary on a hedonic (pleasant–unpleasant) dimension; they have affective or emotional impact, often of devastating magnitude, motivating properties, and similar characteristics.

It is clear that such things as pain and taste have the properties just mentioned, but other sensory events also have them. For example, color probably influences most of us in ways that go far beyond mere discrimination of such things as red and green lights. It would not have occurred to most of us before Bexton, Heron, and Scott (1954)

published their research on perceptual deprivation that a reasonably high flow of information through the senses is essential to the maintenance of adequate psychological functioning. In perceptual isolation, subjects begin to show many indications of serious mental impairment. They even hallucinate. They are also very strongly motivated to get out of the monotonous environment. Indeed, perceptual processes are more than merely informative.

Since a variety of research areas in perception will be discussed later, I need not list perceptual phenomena in order to establish the value of studying perception. I will let my case rest here, with one final word. Just as the area of learning has had an important impact on other areas of study, the influence of perception has been widespread. Its theories and data have been used by psychologists in other areas to almost as great an extent as have those of learning. Often it is possible to approach a given problem area either from the vantage point of a learning-oriented psychologist or from that of a perception-oriented psychologist. For example, in psychotherapy we have today a very energetic conditioning-therapy movement, but at the same time there is a strong gestalt-therapy movement that takes perceptual phenomena as its starting point. In general, it would be impossible to have a firm grasp on contemporary psychology without a reasonable understanding of both perception and learning.

APPROACHES TO THE STUDY OF PERCEPTION

A number of major strategies have been employed by investigators in the area of perception, and it is worth our while to attempt a rough classification of them:

1. Classical introspectionism
2. Phenomenal description
3. Inference-to-structure
4. Physiological approach
5. Inference-from-structure
6. Functional-holistic approach
7. Synthetic approach

Classical introspectionism

Historically the earliest approach to perception (indeed, to experimental psychology as a whole), this involved attempts to analyze the contents of experience into elemental sensations. Just as chemists had analyzed matter into a limited set of elements, psychologists were to look into the contents of their minds and attempt to determine which experiences were elementary. Once the laws of combining these elements were discovered, it was hoped that consciousness could be

explained. Thus the whole gamut of tastes might be analyzed into combinations of such elemental sensations as "sweet," "sour" and "salty." This method, which was purely psychological, worked hand-in-glove with the method of inference-to-structure.

Phenomenal description

This method is similar to classical introspection in that it involves careful description of the contents of the mind. It differs from the classical method because phenomenal description does not involve any attempt to break experience down into elements. Instead, experiences are to be described as directly as possible, without analyzing them. All preconceptions about reality or about theoretical notions of how to categorize experience are suspended. The experience is described, insofar as this is possible, exactly as it occurs to the experiencing individual. Such descriptions *should be* "naive." Even if I am having an hallucination I describe it as I experience it. It may not be real, but it is my phenomenal experience nevertheless.

Inference-to-structure

The inference-to-structure approach has seen heavy use, not only in perception but also throughout psychology. It entails the description of behavior by one of the many possible techniques available, such as phenomenal description or classical psychophysics. After the psychological data have been obtained, speculative anatomical or physiological entities are used to account for the observed data. Note that the anatomical or physiological entities are not *observed.* They are postulated as explanatory constructs to account for what is observed. (If they are actually observed, a transition has been made to what we term the physiological approach.)

A good example of the use of the inference-to-structure approach can be found in the development of classical color vision theory. The Young-Helmholtz trichromatic theory originated in observations of color-mixing phenomena (Helmholtz, 1852b). An observer who is given a color sample and then is provided with pigments of three primary colors, which can be mixed in any proportion, can match the sample color. The presence of three pigments in the retina was postulated to account for the data. The classical antagonist to the three-pigment theory—Hering's (1874) opponent-process theory—started from the observation that colors occur in complementary pairs; once again this is an observation about how observers react to certain stimuli. When certain pairs of colors such as red and green are mixed together, they tend to neutralize each other, appearing gray. In accord with the inference-to-structure technique, Hering postulated the presence of opponent processes in the visual system, which behaved in such a way as to produce complementary color-mixing effects.

Physiological approach

The physiological approach resembles the inference-to-structure method, and the latter is often followed by direct physiological investigation. The critical difference between the two approaches lies in the physiological strategy, which uses direct observation instead of inference. In recent years, for example, direct measures of the color-absorbing pigments in the eye have made it possible largely to resolve the classic disputes about the nature of the physiological process underlying color vision. There are in fact three distinct peaks of pigment sensitivity. The techniques are described later in the section on color vision. For now it is enough to point out that they exist and that they have made it possible to move from one strategy to another, with very gratifying results.

Inference-from-structure

The approach of inference-from-structure could also be called the "anatomical" approach because it places great emphasis on anatomical observation. It is the reverse of inference-*to*-structure. The investigator looks at anatomy and speculates on its function. For example, students of anatomy decided that a very large part of the brain, forming a kind of loop around its central core, was dedicated to the sense of smell. They made this judgment because the olfactory bulbs are so clearly connected to the nose, and they clearly feed into the brain regions in question. However, the anatomists were wrong about the functions of these brain regions. They seem to be more involved with such functions as mating, eating, and fighting. The closeness of their connections to the olfactory bulbs is probably due to some relationship, ancient or contemporary, between smell and their actual functions.

Functional-holistic approach

Functionalists and holists are two different groups of psychologists who would think of themselves, for the most part, as quite opposed to each other. By a functional approach, I mean one that emphasizes the importance of the consequences of perception. Perception is not just sensory. It varies with consequences. The amount of pain a person feels when injured depends on whether the pain means escape from a fearful situation or means loss of a wanted situation. The sensitivity people show when detecting stimuli depends on how well they are reinforced for the detections. People who emphasize this aspect of variation are using a functional approach.

The functional approach suggests that our motivations, expectations, and the like have a lot to do with what we perceive. The holistic approach says that the whole person perceives. Our attitudes, motivations, expectations, and so on help to determine our perception. You can see why I put the two approaches together.

A number of studies show that we have a basic characteristic, called *field dependence* or *field independence,* that influences our perception in a whole variety of situations. This was first discovered when people were placed in a darkened room and asked to move an illuminated rod until it was exactly vertical. Surrounding the rod was a frame that the experimenter tilted to varying degrees. This is called the *rod and frame test.* People differed in the extent to which the tilt of the frame influenced their judgments of the verticality of the rod. Those who were heavily influenced were called field-dependent; those who were influenced little if at all were called field-independent. Other tests, such as one involving the ability to recognize a figure embedded within another figure, revealed that field dependence or independence showed up in the same people across tasks. Thus, there appears to be a basic personality characteristic swaying the perceptual response.

Synthetic approach

The synthetic approach is very different from the ones already mentioned. Science is said to proceed by analysis and synthesis; it breaks things down into their parts and it permits us to create phenomena artificially by synthesizing them from parts. The synthetic approach entails putting together a perceptual phenomenon artificially.

For example, there has long been dispute over how the receiving structures of the inner ear determine the pitch of a sound. One point of view has been that the particular location on the membrane containing the receptors (the basilar membrane) indicates the pitch. George von Bekesy, who won a Nobel prize for his work on the mechanism of hearing, believed that the location was indicated by a "traveling wave." A traveling wave is familiar to anyone who has ever taken a long rope and snapped it up and down so that waves traveled down its length. Similar traveling waves might go down the basilar membrane. Bekesy actually showed that they do by direct observation (*that* was an example of the physiological approach). But, since the wave traveled over the whole membrane, how could it be said to act at a specific location on the basilar membrane? Bekesy argued that the ear could identify the specific location on the membrane where the *peak* activity took place. The ear would have to treat the peak of the distribution of energy as though it were the only region of the basilar membrane activated. But how could Bekesy show that the ear had such capabilities?

One method he used took advantage of the similarities between hearing and skin sensation (Bekesy, 1959). Embryologically, skin sensation is much like hearing. The basilar membrane might even be considered a very condensed bit of skin, in the sense that there are many more receptors per unit area on it. Sometimes sounds can even be confused with taps on the skin! So Bekesy built a vibrator to fit on the

Key Ideas Box 8.1: Approaches to the study of perception

There are several fundamentally different approaches to the study of perception.

1. *Classical introspectionism* involves the analysis of mental contents, seeking the elements of consciousness and the laws of their combination.
2. The *phenomenological method* emphasizes description of experience without analysis, criticism, or theoretical preconceptions.
3. *Inference-to-structure* is a method whereby inner structures are hypothesized as a result of behavioral observations.
4. The *physiological approach* involves direct measurement of activity in bodily structures that are supposed to underly perception.
5. With *inference-from-structure* anatomy and physiology are observed directly and the psychological functions are hypothesized to be mediated by the structures.
6. The *functional-holistic approach* emphasizes multiple determination of perception instead of focusing solely on receptive structures. It turns attention to motives, expectations, and the like.
7. The *synthetic approach* entails reconstructing systems that duplicate the functions of the perceptual system under study. As in other areas of science, the ability to synthesize something demonstrates the validity of our understanding.

arm, and he vibrated it at various rates, paralleling the varying rates of vibration of the inner ear produced by sounds of varying frequency. He found that, even though the entire arm was vibrating, subjects reported feeling activity at a particular spot on the arm, and the spot changed when the frequency of the vibration changed. Thus a system much like the basilar membrane recognizes a traveling wave as activity on a particular location of the receptive surface.

Psychophysics

Psychophysics occupies a very important place in the history of science as a whole and of psychology in particular. Centuries were required for science to establish itself as a truly fruitful way of coming to understand phenomena. Scientists had to fight hard every inch of the way in order to convince the world that traditional views should be replaced by scientific ones. Today we can hardly imagine the fierce antagonism with which discoveries of the early physical sciences were greeted several centuries ago, even though we have heard of famous

incidents such as the inquisition of Galileo. The scientific view of the physical universe prevailed over unscientific views, and its antagonists were forced to make their peace with it. But they rejected it stubbornly at every point where they could possibly take a stand. When they had no choice but to accept the scientific approach to physical nature, many individuals argued that scientific principles found in physical nature did not apply to living systems. This was the point of view known as *vitalism*—that living things had a special life force that differed from ordinary physical forces and that made them living things.

The great battle against vitalism, though the war is even now not entirely won, took place in the nineteenth century when it became clear that the law of conservation of energy (that energy is neither created nor destroyed in any chemical change) applied to living creatures as well as to inanimate matter. Living things could be studied as physical systems, since that is just what they are. The impact of such findings on our present world has been tremendous. The encouragement provided by such early demonstrations has led to the great discoveries that have taken science deep into the secrets of living nature—so deep that we now understand a very great deal about the chemistry of genetic structure underlying the construction of a living organism.

Almost in parallel with the attack on vitalism was the Darwinian analysis of the origin of species, which implied that the special place of man in nature was not so special after all. Man, as were other organisms, was the end result of a natural selection process in which organismic characteristics that were well adapted to an environment led to survival and therefore to perpetuation of the genes specifying them. Living things could be brought under the sway of natural science, and man was no exception.

Yet, even here, it was possible to separate from science some certain aspects of the human organism. To be sure, the body had evolved from lower forms through natural selection, and the evidence for this quickly became overwhelming. But the mind of man was something not found in the lower animals. It emerged suddenly with the emergence of man. Furthermore, though it was possible to measure physical things, mental things were not capable of being measured. Without measurement, there can be no science, and the mind of man was therefore outside the domain of natural science.

It was the early psychophysicists who, sometimes unwittingly, broke down the latter argument. Instead of arguing whether mental events could be measured, they simply began measuring one class of mental processes, the sensory or perceptual ones. The kinds of things they measured were at first very simple, such things as judgments of the weight of objects and of changes in their weight, but the important point was that psychological events *could be measured.* There is an old

saying that the beginning is more than half of the whole, and even a small beginning is, after all, a beginning.

The beginning of psychophysics in the late nineteenth century was the beginning of psychology as an experimental science. There were still many places for resisters to put up their blockades. One could, for example, hold that measurement could be accomplished only with such lower mental processes as sensation and that higher mental processes, such as memory and intelligence, were unmeasurable. But it was not long after the great psychophysicist Fechner published his *Elemente der Psychophysik* (1860; translated 1966) that Ebbinghaus (1885), inspired by Fechner's work, went on to measure memory and forgetting. And soon after Ebbinghaus's experiments, Binet went on to apply psychophysical methods to the measurement of intelligence (Binet & Simon, 1916). During the late nineteenth and early twentieth centuries psychologists found ways to measure a very impressive range of mental events, including the length of time taken in making simple choices, judgments, and associations. By the early twentieth century it was clear that mind could be included in the domain of natural science. Thus, all scientists owe psychophysics a great debt.

PSYCHOPHYSICS AS A MODEL OF PSYCHOLOGICAL METHOD

The heart of experimental psychology is the search for functional relationships. We seek the relationship between such things as the physical intensity of a tone and the frequency of a response ("I hear it") or the number of marbles a child gets and how fast it emits button pushes that have marbles as their consequences, or the relationship between the rates of emission of expressions of opinion when they are followed by a friendly nod or when they are not.

Could we have added to the preceding examples that psychologists try to find relationships between such things as kissing and love? Kissing is a reasonably identifiable event, but one hardly expects to find an experimental psychologist dealing with "love" as a variable. Yet, in fact, whether "love" can be treated as a scientific variable depends only on whether it can be measured. If the meaning of the word "love" can be made sufficiently explicit so that consensual validation can be attained about its presence or absence, or, better yet, so that a reliable quantitative *scale* of it can be generated, there is nothing to prevent its being studied as a variable just like any other variable.

In order to find a functional relationship, it is necessary to devise a method of measuring the events to be related. Psychophysics deals with the problems entailed in detecting, recognizing, discriminating, and scaling events (Galanter, 1962); a knowledge of its principles is of

great value in learning how to cope with the problems of measurement. Psychophysics is, in a sense, the science of observation, and the importance of observation to science in general has been documented earlier.

Psychophysics is also important because (1) it is the oldest area of scientific psychology, and perhaps the most advanced; and (2) it encompasses most of the main problems of psychology. The problems of threshold measurement may seem remote from the kinds of problems most psychologists wish to solve, but a closer inspection shows that psychophysical methods and psychophysical problems are encountered everywhere. For example, the creation of IQ tests has many facets that are formally identical to aspects of threshold measurement. Attempts to derive IQ tests that are free from the influence of environmental variables, that reflect true "biological capacity," are analogous to attempts to find the "true threshold," the threshold that reflects the capacity of receptor organs. Even humdrum problems such as those faced by a teacher who wishes to correct for guessing on a multiple-choice test are like psychophysical problems.

In sum, the most important conceptual problems of psychology can be found in psychophysics. And it is worth adding that the most important problems of research strategy and tactics are encountered there as well. A person who understands psychophysics thoroughly is already a psychologist.

To repeat, there are four basic problems in psychophysics:

1. The detection problem—the problem of detecting the presence of an event
2. The recognition problem—the problem of identifying an event
3. The discrimination problem—the problem of telling whether some event is different from a standard
4. The scaling problem—the problem of developing some sort of "yardstick" for measuring psychological events

We will deal with three of these—detection, discrimination, and scaling. Although the recognition problem is important and interesting, it is a bit complicated for present purposes (see Hake & Rodwan, 1966, for information concerning it).

DETECTION

The detection problem is basically this: How much of a physical stimulus does it take before people can detect it? The fact is, there are many physical energies we cannot detect. Figure 8.2 shows that only a small part of the electromagnetic spectrum is visible. Everybody knows that dogs can hear a whistle too high in frequency for the

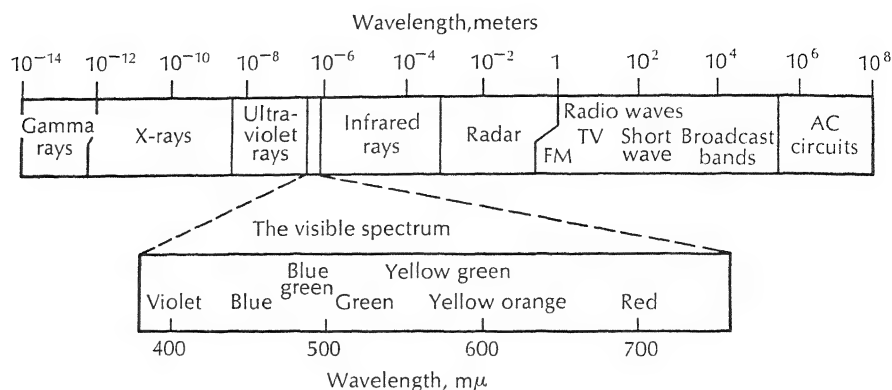



FIGURE 8.2 The total electromagnetic spectrum. That portion of it normally affecting the eye, occupying a relatively narrow band of wavelengths, has been enlarged in the lower part of the diagram. (From Chapanis, Garner, & Morgan, 1949.)

human ear. Bats and porpoises do even better than that. Even children hear higher pitches than most adults. The question is how to measure the lowest point at which physical stimuli are detectable for a given organism.

For one thing, no single point will emerge from direct measurements of sensitivity. There is going to be error of measurement, noise in the background, or what have you. Those background events get confused with the signal to be detected. Furthermore, there are variations within the person who is trying to detect the stimulus. From day to day and from moment to moment alertness and energy may wax and wane, influencing the point of detection.

So we get back to the old rationale for using averages. No single measure is reliable. Observations in this error-filled world are always combinations of the true measure and error of measurement. Fortunately, errors of measurement tend to be above the true measure as often as they are below it, so we can get at the true measure by taking some kind of average of the real observations.

Determining detection thresholds

The crossover between the place where stimuli are not detectable and where they are is called the ***threshold.***[1] Figure 8.3 illustrates the ideal notion of threshold, in which there is an abrupt change from no

[1]The Latin word for threshold is *limen,* and early psychophysicists used the Latin form. It is still encountered and should be recognized as an equivalent for "threshold." In place of the word "detection" we sometimes hear "Reiz," "stimulus," or "absolute." "Stimulus threshold," "absolute threshold," detection threshold," "stimulus limen," "absolute limen," and "Reiz limen," all mean the same thing. The remaining possible combinations of noun and adjective are, for purely historical reasons, not generally used.

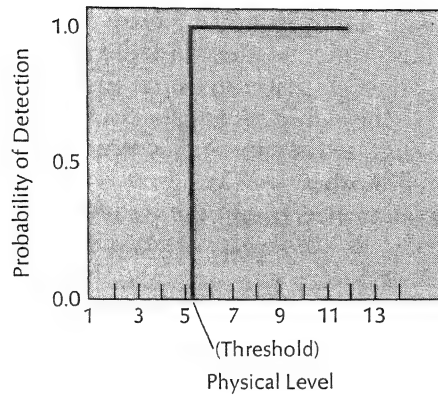



FIGURE 8.3 Idealized graph of detection threshold showing the probability of detection jumping from zero to 1.00 when a critical (threshold) magnitude of the physical stimulus is reached.

detection to perfect detection. Contrast this to Figure 8.4, which shows something closer to the true state of affairs. The curve is gradual instead of abrupt because various sizes of error have been present at different times, obscuring the true function.

If there is no discrete point that divides the area of no sensation from the area of sensation, what is the threshold? How do we ever determine it? The traditional approach follows this course: Assume that there is a "true" threshold. Assume further that the reason we fail to find it in our direct measurements is that there are sources of "error" (errors of

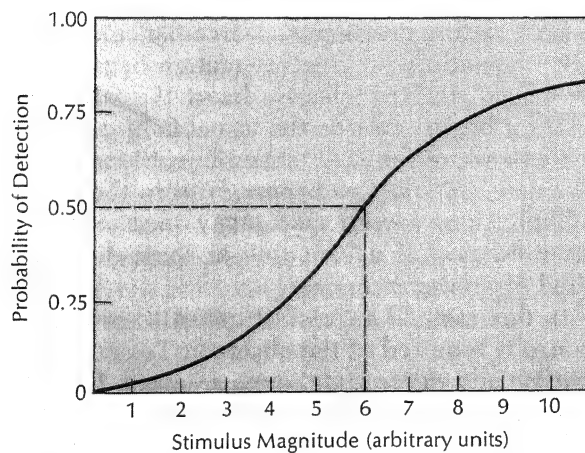



FIGURE 8.4 Idealized graph of detection threshold illustrating continuous variations in the probability of detection with physical stimulus magnitude. The detection threshold is the physical magnitude corresponding to a detection probability of 0.5.

measurement, fluctuations in "noise" level, and other factors). Assume also that the errors are distributed evenly above and below the true threshold. It makes sense to estimate our true threshold by taking the midpoint of the curve of actual measurements. By the midpoint we mean the point on the physical continuum that results in a probability of "yes" that is equal to 0.5 or, in other words, where the frequency of "yes" responses is equal to the frequency of "no" responses. *This 50 percent point is what psychophysicists in actual practice call the detection threshold.*

There are a number of methods for determining thresholds. They are called simply *the psychophysical methods.* We will go into a number of them.

The method of limits (minimal change)

We begin with the method of limits (also called the *method of minimal change*). Suppose you have noticed some symptoms that lead you to suspect that you might have a neurological problem. Full of fear, you go to a neurologist, who takes you through a variety of simple tests—a kind of preliminary neurological examination designed to determine whether more elaborate tests are needed. One thing neurologists are likely to do is ask the patients to look straight ahead while they move their fingers in the periphery of the patient's vision. At first they will place their fingers outside the patient's field of vision and ask whether motion can be seen; then they will move inward to within the patient's field of view. This is repeated in several parts of the visual field. The technique used by the neurologist is a very crude form of the method of limits. The neurologist has not controlled for the fact that you might be able to hear the motion of fingers, might be able to see body motions that tell you that the fingers are moving, and that there might be order effects, depending on whether motion begins from within the visual field where you can clearly detect the stimulus and moves out or whether it begins outside the visual field where you cannot detect the stimulus and moves in. Furthermore, the neurologist is likely to take only a few measures and place faith in their reliability, whereas the psychophysicist would take many measures and place faith in an average figure. Let us examine in some detail the psychophysicist's method of minimal change.

With this method a "yes" (I detect it) or a "no" (I do not detect it) response is required of the observer. To control for order effects, both *ascending* and *descending* series are used. In an ascending series, the experimenter begins with a physical stimulus value well below that which can be detected and increases it in small steps. A descending series begins with a value well above the threshold and works down. These two types of series are employed because the "threshold" value obtained with an ascending series will commonly be different from the

value obtained from a descending series. In other words, there is typically an order effect.

The number of steps (or the starting point) in ascending and descending series can be varied so that the observer cannot be falsely consistent by simply counting the number of trials to a response change. One might, for example, notice that one changed responses at the eighth presentation for the first series and thereafter simply change on Trial 8, regardless of one's awareness of the stimulus.

Differences on ascending versus descending series are due to *constant errors.* If an observer adopts a *criterion* or *decision rule* demanding no change of response until the experience overwhelmingly indicates that the stimulus is present, the result is an *error of habituation.* If the rule determines change of response when the experience merely tends to indicate the presence of the stimulus, the result is an *error of anticipation.* If we are willing to assume that these "errors" occur equally often on ascending and descending series, their effects can be averaged out after using both kinds of series. Thus the method of limits employs the familiar procedure of counterbalancing in order to control for order effects.

As I said earlier, we must take the average of many readings of the threshold in order to get a good estimate of the true threshold. By taking several measures, the experimenter can get reliable results and can also control for order effects. The experimenter obtains a threshold estimate, T, for each of the series. Any one of the T values would be inadequate as a measure by itself, because, among other things, it has a confounded effect (a given series must be either ascending or descending, not both). T is therefore a crude, confounded threshold estimate. Specifically, T is equal to half the distance between the physical value *before* the observer's detection response changed (on ascending series, from "no" to "yes"; on descending series, from "yes" to "no") and the physical value *at which* the response changed. Thus, if the response changed at a physical value of 11 (arbitrary units) and the value just preceding 11 was 10, then $T = 10.5$. Psychophysicists use the average of many T values as the threshold value. The arithmetic mean is generally used, that is, the threshold is obtained by adding up all T values for all ascending and descending series and dividing the resulting sum (ΣT_i) by the total number of series *(N)* employed. The detection threshold is expressed in a formula as

$$T = \frac{\Sigma T_i}{N}$$

where $T_i = T_1, T_2, \ldots, T_n$ and $N =$ the total number of T values.

A worked example of an experiment using the method of limits to obtain the detection threshold is given in Table 8.1. The data are imaginary.

TABLE 8.1 Determination of Stimulus Threshold by the Method of Limits

Stimulus,	Responses					
Hz	*Trial 1*	*Trial 2*	*Trial 3*	*Trial 4*	*Trial 5*	*Trial 6*
25	—	Y				
24	—	Y				
23	—	Y	—	—	Y	
22	—	Y	—	—	Y	
21	Yes	Y	—	—	Y	
20	Y	Y	—	—	Y	
19	Y	Y	—	—	Y	
18	Y	Y	—	—	Y	
17	Y	Y	—	—	Y	
16	Y	Y	Y	—	Y	
15	Y	?	N	Y	N	Y
14	No	—	N	N	—	N
13	—	—	N	N	—	N
12	—	—	N	N		
11	—	—	N	N		
10	—	—	—	N		
9	—	—	—	N		
8	—	—	—	N		
7	—	—	—	N		
6	—					
5	—					
4	—					
3	—					
2	—					
1	—					
T =	14.5	15.5	15.5	14.5	15.5	14.5

The method of adjustment (average error)

With the adjustment method the detection threshold is determined by giving control of the physical stimulus to the observer, who makes continuous adjustments of the stimulus. It is much like the procedure we all use for tuning in a radio. The observer is usually permitted to hunt back and forth until satisfied that the "correct" threshold value has been reached. This differs from the method of minimal change, since with the latter method the experimenter does not move the stimulus values back and forth in a range near the threshold, but instead gives ascending or descending series and simply records the first point of the observer's change of judgment. Something like ascending and descending series can be employed with the method of adjustment by starting sometimes with values above the threshold and at other times with values below the threshold. However, since the observer can oscillate above and below the "threshold" area, this is not exactly like using ascending and descending series with the method of minimal change.

How do we determine the detection threshold with the method of adjustment? We simply take the mean of "T" values determined over several trials. Hence the detection-threshold formula is the same as that for the method of minimal change.

The method of constant stimuli

With the method of constant stimuli, we use a fixed set of physical stimuli and find out how often each is detected. These stimuli must be around the threshold value, so you often have to do a crude threshold measure with one of the other methods before beginning with the constant-stimulus method. Customarily we present stimuli in random order rather than in ascending and descending series. There are many ways of getting the detection threshold from the resulting data. If subjects respond to one of the stimuli 50 percent of the time, that stimulus value is the threshold. Usually, values near 50 percent will occur; then you have to get the threshold by some indirect method.

One very simple indirect method is to plot the obtained values on a graph, with the probability of a "yes" response on the vertical axis and the physical value on the horizontal axis. Next draw a freehand S-shaped curve that fits as nearly as possible to the data. Next project a horizontal line from the 0.5 point on the vertical axis until it intersects the S-shaped curve. Finally, drop a vertical line from the intersect to the horizontal axis, and read the physical value. This is a crude technique, but it is reasonably accurate.

A similar result (but often not the same) can be obtained numerically by "linear interpolation." For example, if the values closest to 50 percent are 40 percent "yes" responses for a physical value of 10 (arbitrary units) and 55 percent for a physical value of 20, the physical value corresponding to the 50 percent point must lie somewhere between 10 and 20. Since 50 percent is two-thirds of the distance from 40 to 55, that is,

$$\frac{(50-40)}{(55-40)}=\frac{2}{3}$$

we take two-thirds of the difference between the two physical values in question and add it to the smaller value in order to get the detection threshold. Hence,

$$(20-10)\times\frac{2}{3}=6.7 \quad \text{and} \quad 10.0+6.7=16.7=T$$

Other methods for deriving the threshold from data derived by the constant-stimulus method may be found in Woodworth and Schlosberg (1953).

different another stimulus, called a *comparison stimulus*, must be in order to be identified as different from the standard.

For example, you might start with a stimulus weighing 100 grams and ask how much weight change is necessary to make a stimulus just noticeably different from 100 grams. Both the comparison and the standard stimulus are clearly detectable, and the problem is to find out when the comparison stimulus seems different from the standard.

The same three methods used to get detection thresholds can be modified to give discrimination thresholds.

Obtaining the discrimination threshold by the method of limits

Let's begin by focusing attention on the goal, the unknown we want to find. For now the subleties will be left out and we will assume something of an ideal world. The discrimination threshold is the smallest amount of physical change giving rise to a sensation different from that produced by the standard. Consider that there must be some mean level at which the comparison and the standard will appear subjectively to be equal. This is called the *point of subjective equality* (PSE). It is not always, or even usually, equal to the point of physical equality of comparison and standard. Moving from an observer's mean point of subjective equality (either up or down) we will eventually reach a point where the comparison is experienced as different from the standard, and that point will be the *just noticeable difference* (jnd) or *difference threshold.*

You can see that points of subjective equality are in a region in which the comparison and the standard are confused by the observer. To say that they are confused is the same as to say that they appear to be equal, that the observer cannot tell them apart. Furthermore, comparison and standard continue to be confused until they are detectably different, or, in other words, all the way from the point of subjective equality to the point of the just noticeable difference. Keep in mind that this confusion applies whether the comparison is greater than or less than the standard. The region of confusion ranging from the lower cutoff through the point of subjective equality and on to the upper cutoff is called the *interval of uncertainty* (IU). The point of subjective equality divides the interval of uncertainty into two halves, each of which is exactly one just noticeable difference. Hence we can find the discrimination threshold by first finding the range of values between the upper and lower points at which the comparison is perceived as different from the standard (the interval of uncertainty) and dividing by two. Stated symbolically:

$$DL = \frac{IU}{2}$$

where DL = the difference limen (threshold) and IU = the interval of uncertainty.

You can see that the problem of determining the discrimination threshold is fundamentally a problem of finding the interval of uncertainty, which will then be divided by two. Since the interval of uncertainty is the range of values between the point where the comparison stimulus is judged greater than the standard and the point where it is judged less than the standard, our task is to find these upper and lower cutoff values.

With the method of limits, the experimenter takes a comparison stimulus value clearly below the standard, and moves up in small steps from the place where the observer consistently judges the comparison to be "less than" the standard, through the region where the comparison and standard are confused and finally on to the place where the comparison is judged "greater than" the standard. By continuing each series all the way through the −, = or ?, and + values, a measure of both the lower and the upper cutoff is taken on each series. Of course, it is necessary to counterbalance for order of presentation, so the experimenter must use both ascending and descending series. Furthermore, no single cutoff measure will give a true value by itself. It is necessary to run the observer through several different ascending and descending series in order to get the average of several upper cutoff values (T^+) and several lower cutoff values (T^-).

The difference between the mean of the T^+ values and the mean of the T^- values is the interval of uncertainty, and one-half of it is DL. Thus, if there were five T^+ and five T^- values, DL would be

$$\mathrm{DL} = \frac{\dfrac{T_1^+ + T_2^+ + T_3^+ + T_4^+ + T_5^+}{5} - \dfrac{T_1^- + T_2^- + T_3^- + T_4^- + T_5^-}{5}}{2}$$

OBTAINING THE DISCRIMINATION THRESHOLD BY THE METHOD OF LIMITS: AN EXERCISE AND A WORKED EXAMPLE The best way to learn about psychophysical methods is to measure a threshold yourself. But the human senses are so remarkably sharp that most threshold measures require precise measuring equipment. Still, there are some cases in which fairly accurate measures can be gotten from equipment available in the average home.

The task is to measure the difference threshold for thickness as judged by pinching a stimulus between thumb and forefinger. It is important not to look at the stimuli as you make judgments. Stimulus materials are made by clipping together various numbers of sheets of paper. The pinching is done on the corner of a given paper stack, very near the clip. In this way the sheets are held firmly together.

I will describe an experiment I conducted at home. I carefully counted out stacks of paper sheets ranging from 2 to 16 sheets. Several times I found that I had miscounted, so it is important to count very carefully, and at least twice. A small error will have an effect on your

results because your sense of touch is remarkably accurate. The stacks were clipped near one corner, and the sheets arranged in order from smallest to largest on a large table with the corner nearest the clip readily available, hanging over the table's edge. Ascending and descending series were alternated, and the time interval between judgments held fairly constant. A data sheet was prepared ahead of time with "A D A D A D A D A D" indicating ascending and descending series at the top and the numbers of sheets from 16 to 2 (top to bottom) down the left side of the sheet. Entries were made on the data sheet for each judgment. The symbols used were "+" for "greater than"; "−" for "less than"; "=" for "equal to"; and "?" for "doubtful."

Classical psychophysicists emphasized the importance of using trained observers because there is a tendency to become less erratic and more sensitive with increasing experience at making psychophysical judgments. You will find that results will be unsatisfactory unless you give yourself a number of preliminary runs at making judgments. Eventually you will notice the results becoming somewhat stable, and then you are ready to gather data.

Results from my experiment are given in Table 8.2. The standard

TABLE 8.2 Results of Experiment Using Method of Limits to Determine Difference Threshold

Number of Sheets	A	D	A	D	A	D	A	D	A	D
16		+		+		+		+		+
15		+		+		+		+		+
14		+		+		+		+		+
13		+		+		+		+		+
12		+		+		+		+		+
11		+		+		+	+	+		+
10		?		+		+	?	?	+	?
9		=		=	+	?	?	=	?	=
8	+	=	+	=	=	=	?	=	?	=
7	=	=	=	=	=	=	=	=	=	=
6	−	−	=	−	=	=	?	−	=	=
5	−		−		−	−	−		−	=
4	−		−		−		−		−	−
3	−		−		−		−		−	
2	−		−		−		−		−	
T^+	7.5	10.5	7.5	9.5	7.5	9.5	10.5	10.5	9.5	10.5
T^-	6.5	6.5	5.5	6.5	5.5	5.5	5.5	6.5	5.5	4.5

$$\Sigma T_i^+ = 93 \quad \bar{T}^+ = 9.3 \quad \text{IU} = 9.3 - 5.8 = 3.5$$

$$\Sigma T_i^- = 58 \quad \bar{T}^- = 5.8 \quad \text{DL} = \frac{\text{IU}}{2} = \frac{3.5}{2} = 1.85$$

$$\text{PSE} = 5.8 + 1.8 = 7.6$$

Tactile judgments of thickness were made by pinching varying numbers of sheets of paper at the edge between thumb and forefinger. A = ascending series, D = descending series. Numbers in the far left column indicate number of sheets in a given stack. Sheet thicknesses ranged approximately from 0.003 to 0.005 in. The standard stimulus was 7 sheets thick.

stimulus was arbitrarily set at 7 sheets and the difference threshold was 1.8 sheets. Measurements made with a micrometer on various numbers of sheets indicated that a sheet was from about 0.003 to 0.005 in thick. Thus, the measurements give a pretty good idea of the actual threshold. However, the main value of the experiment is to gain experience with the method.[2] Even if you are short of time, it is worthwhile to run through one or two ascending and descending series. The discussion of psychophysical methods will be much easier to understand if you do so. If, on the other hand, you enjoy doing such experiments and have enough time, you can go on and get the difference threshold with the method of constant stimuli, and you can also attempt to verify the constancy of Weber's fraction, which we will discuss soon.

Incidentally, if you want to be more exact in your measures of this threshold you can use wires or (what amounts to the same thing) nails of various thicknesses for stimuli and measure them with a micrometer.

Obtaining the discrimination threshold by the method of constant stimuli

With the method of constant stimuli, thresholds must be obtained by deducing them from the relative *frequencies* of judgments of "greater," "equal or doubtful," and "less." Begin with a standard stimulus and a set of comparison stimuli ranging on either side of it that will encompass the threshold. The experimenter must have some rough idea of the threshold value in order to select the stimuli appropriately, so a preliminary determination, perhaps with the method of limits, is valuable.

It is typical to select about seven comparison stimuli, and each of these is repeatedly paired in random order with the standard. For example, each comparison stimulus might be presented with the standard for ten times, but interspersed at random with the other comparison stimuli. Each comparison stimulus will yield its own frequency of each of the three possible response categories ("Greater," "Equal or doubtful," and "Less").

Assume that we have gotten the response frequencies for each of the comparison stimuli. How do we get the interval of uncertainty, and then the difference threshold? Once again, we look for the upper cutoff where the stimulus is just perceptibly greater than the standard and the lower cutoff where it is just perceptibly less. Look at Figure 8.5. In that figure the judgments of "less" have been plotted in the top curve and

[2]You may prefer to work with lengths of lines, as the great psychophysicist Weber did. My own students do this and reproduce his results very closely. They carefully draw the lines on index cards.

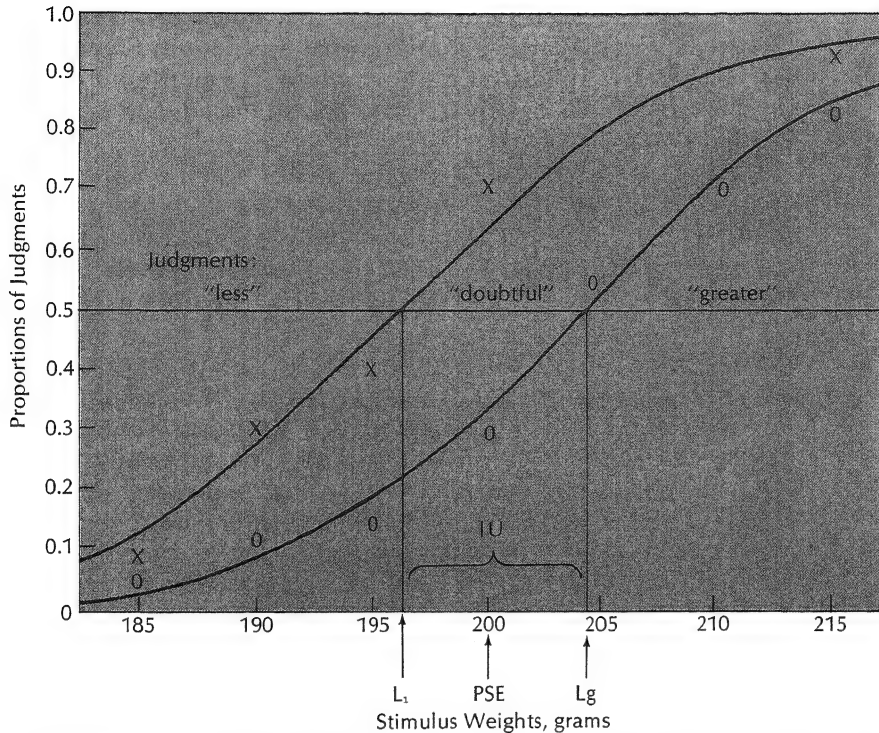



FIGURE 8.5 The upper line has been drawn through plotted frequencies (seen as "X"s) for judgments of "less." The lower line has similarly been drawn through plotted frequencies (seen as "O"s) for judgments of "greater." A line has been projected from the 50 percent point (proportion of judgments = 0.5) of the line for "less" judgments to the horizontal axis and from the line for "greater" judgments to the horizontal axis. Where the projected lines strike the horizontal axis we have the lower and upper limits of the interval of uncertainty (IU). The values on this graph are roughly 197 and 204. The middle of the interval of uncertainty is the point of subjective equality (PSE), and one-half of the interval of uncertainty is the difference threshold (DL). (From Guilford, 1954.)

judgments of "greater" have been plotted in the bottom curve. The region of uncertainty lies between the two curves.

We take as the point where the stimulus is just perceptibly greater the physical value where it is judged "greater" half the time. Similarly, we take as the lower cutoff the place where the stimulus is judged "less" half the time. If the observed values fail to include any that give exactly 50 percent of the two types of judgment we want, then we have to estimate where the 50 percent points would have been if we had happened to pick the right comparison-stimulus values. There are several ways to do this, but a simple and accurate one is to plot the observed values and fit a curve to them on a graph like the one in

Figure 8.5, and then to draw a line that intersects the curve at the 50 percent level. Notice that this is what was done in Figure 8.5. There are no *real* data points for the spot where "proportion of judgments" on the vertical axis equals 0.5. But the estimated data will give good results.

Having estimated the 50 percent points for the upper and lower cutoffs, we simply drop a vertical line from each of the two cutoffs down to the horizontal axis (see Figure 8.5) and read the physical values for each of them. The difference between these two values is the interval of uncertainty, and one-half of it is the difference threshold.

A TWO-CATEGORY METHOD OF OBTAINING THE DIFFERENCE THRESHOLD BY THE METHOD OF CONSTANT STIMULI. Psychophysicists have devised a nice approximate method for getting a discrimination threshold by the method of constant stimuli. The experimenter does not allow the observer to make the judgment of "equal or doubtful." It is a forced-choice situation in which the observer must respond either "less" or "greater"; thus the method is called the "Two-Category Method." The subject who does not know whether to choose "less" or "greater" must guess.

The method is essentially like the three-category method. Several comparison stimuli are paired repeatedly and in random order with the standard stimulus, and the relative frequencies of the two response categories are plotted. Since judgments of "doubtful" and "equal" are not permitted, there is no obviously apparent interval of uncertainty. So how do we get the difference threshold?

Psychophysicists have found that the distance between the points where 25 percent and 75 percent of the judgments are in the "greater" category provides a good estimate of the interval of uncertainty. Consequently, we have only to plot the frequency of "greater" judgments on the vertical axis of a graph with the values of the comparison stimuli on the horizontal axis, and follow the same procedure used to get upper and lower cutoffs with the three-category method. That is, project a line down from the 25 percent point to the horizontal axis and another line from the 75 percent point to the horizontal axis. Subtract the lower from the higher physical value obtained and divide by two in order to get the discrimination threshold.

WHY USE THE METHOD OF CONSTANT STIMULI? Since you need to have an idea of the approximate threshold before you can use the method of constant stimuli, why bother with it? A major advantage of this method is that you can give a trial in a very short time. If you want to see what happens to the threshold under the influence of some independent variable that decays quickly over time, this is a very useful property of the method. Many times, experimenters want to

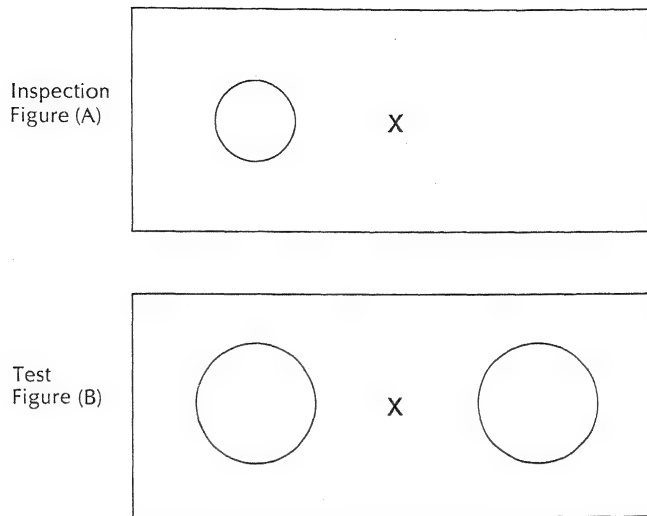



FIGURE 8.6 A figural aftereffect. Inspection of the top figure causes the left circle of the test figure to seem larger (in both cases, fixation is on the X).

know about changes that take place in the threshold when you adapt the subject to some stimulus for varying amounts of time. These adaptation effects may be very short-lived, so the method of constant stimuli is a nice thing to have around.

A good concrete example of the usefulness of the method is the study of *figural aftereffects.* If you look at something for a while (call it an *inspection figure*), then shift to looking at something else (call it a *test figure*), the test figure looks different because of the prior inspection. Figure 8.6 shows the kind of thing I have in mind. Gaze toward a small circle for a while, then shift to a larger one with your gaze centered in the same relative position as to the small one. It makes the larger circle appear even larger than it would normally. The change shows up by contrast to a control circle like the one at the right in the test figure, which is not centered in the same place as the inspection figure. Actually there are a lot of controls you must run in order to get a clean measure of a figural aftereffect, but the main point here is that such effects exist and that *they do not last very long.* At least the big effects go away very quickly.

If you tried to measure these effects using a method of limits or of adjustment, the subject would take a lot of time in an ascending or descending series, or in adjusting the stimulus around the threshold value. During this time, the aftereffect would decay, and you would end up with a very poor measure of it. Since the constant stimulus method can be based on rapid judgments made to stimuli presented quickly, it is ideal for this kind of situation.

Obtaining the discrimination threshold by the method of average error

The two-category method of constant stimuli illustrates an important methodological device used by psychophysicists. With that method there is, strictly speaking, no direct determination of an interval of uncertainty, nor is there a direct determination of the difference threshold. The physical separation between the points where the observer says "greater" 25 percent of the time and 75 percent of the time is a good approximation to the interval of uncertainty, and is at least related to it in an orderly way. Psychophysicists are willing to content themselves with a measure that is at least linearly related to the difference threshold as obtained by the more direct methods.

With the *method of average error,* which is also called the *method of adjustment,* the difference threshold is also obtained indirectly. What the observer does *directly* is to try to set a comparison stimulus so that it is equal to the standard—to directly determine the point of subjective equality, not the point where the comparison stimulus appears different from the standard. Naturally, observers are asked to determine a point of subjective equality repeatedly so that the experimenter can get a stable estimate, an average of individual judgments.

How can the difference threshold be gotten from measures of the point of subjective equality? You must remember that the difference threshold is essentially a measure of confusions. The interval of uncertainty is an interval in which the comparison stimulus is confused with the standard stimulus. As you know, the individual estimates of the point of subjective equality will vary around the mean. The variability provides an index of the degree of confusion. Experimenters often use the familiar standard deviation as an index of the difference threshold. The standard deviation is usually not equal in magnitude to the difference threshold as obtained by the more direct methods, but it is linearly related to it.

So the method of average error is basically very simple. It merely involves asking the observer to adjust a comparison stimulus over a substantial number of trials so that it appears to be equal to a standard stimulus. The observer is usually allowed to adjust the comparison stimulus back and forth in the region of apparent equality until satisfied. After a sizeable number of judgments, their mean is taken to be the point of subjective equality. The difference between the point of subjective equality and the point of physical equality is the constant error. This may be of interest when, for example, the method is used to measure the size of a visual illusion. The standard deviation around the mean can be used as an index of the discrimination threshold.

THE USE OF THE METHOD OF AVERAGE ERROR: AN EXERCISE AND A WORKED EXAMPLE Experimental psychologists

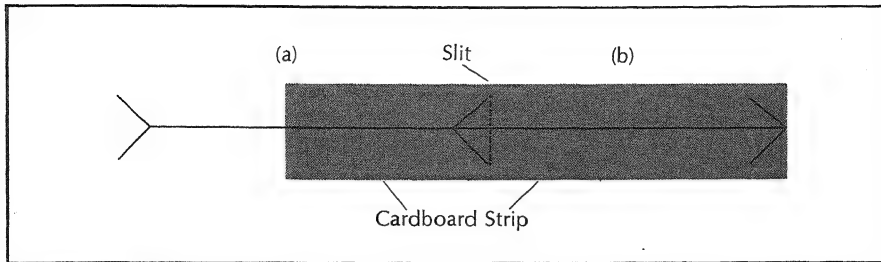



FIGURE 8.7 The Müller-Lyer illusion. Components (a) and (b) are equally long. The apparatus consisted of a large piece of cardboard on which a straight line with "feathers" was drawn (left of figure). A smaller strip of cardboard with an arrow drawn on it was inserted through a slit in the larger piece. By sliding the smaller piece with the arrowhead back and forth, the length of the arrow could be varied over a wide range until the part with the feathers and the part with the arrowheads appeared equal.

often measure illusions by the method of average error. An illusion is simply a constant error—an error that occurs consistently in a given direction. Here we will use the method of average error to measure the extent of the Müller-Lyer illusion. This is the error in judging the length of lines that occurs when "arrowheads" versus "feathers" are put at the ends of the lines (see Figure 8.7). Observers generally tend to underestimate the length of a line when the "feathers" are at its ends.

You may have an apparatus available for measuring the illusion, but, if not, you can construct one of cardboard in a few minutes. Figure 8.7 shows one method of construction.

Many different variables might influence the measurements you will make. You should try to use the methods of dealing with confounded variables that you have learned in order to control for such influences. There are two obvious factors of importance. The first is whether the comparison stimulus starts larger than ("out") or smaller than ("in") the standard. We will alternate "out" and "in" trials. The second is whether the comparison stimulus is to the right or to the left of the standard. Alternate this as your apparatus permits.

On each trial, the observer is presented with the comparison and the standard, the comparison being set at a level either obviously larger or obviously smaller than the standard. The observer is then asked to adjust the comparison until it appears equal to the standard. The length of the standard is measured before the experiment starts, and the length of the comparison stimulus is measured and recorded after each judgment. Table 8.3 shows how to set up an appropriate data sheet, the "outs" and "ins" being indicated down the far left column and the right or left position of the comparison stimulus being indicated at the top of the sheet.

Table 8.3 gives the results of an actual experiment that I did with a home-made cardboard apparatus. Measurements were made in units of

TABLE 8.3 Results of an Experiment to Determine the Magnitude of the Müller-Lyer Illusion.

Comparison	Position of Comparison Stimulus											
"In"	*Right*			*Left*			*Right*			*Left*		
"Out"	*X*	*d*	*d*²	*X*	*d*	*d*²	*X*	*d*	*d*²	*X*	*d*	*d*²
O	74	2	4	70	2	4	71	1	1	72	0	0
I	69	3	9	67	5	25	69	3	9	73	1	1
O	73	1	1	73	1	1	72	0	0	72	0	0
I	71	1	1	69	3	9	69	3	9	73	1	1
O	76	4	16	74	2	4	72	0	0	73	1	1
I	72	0	0	70	2	4	70	2	4	74	2	4
O	75	3	9	76	4	16	77	5	25	77	5	25
I	74	2	4	70	2	4	69	3	9	69	3	9
O	76	4	16	73	1	1	73	1	1	77	5	25
I	73	1	1	71	1	1	71	1	1	72	0	0

"In" and "Out" indicate whether the comparison started out noticeably larger or smaller than the standard. *X* = score measured in units of one-sixteenth of an inch; *d* = deviation from the mean of the scores; d^2 = square of the deviation of that score from the mean of the scores.
Standard = 80 units
Mean of judgments = point of subjective equality = 72 units
Constant error = magnitude of illusion = (standard) − (point of subjective equality) = 8
Standard deviation =

$$\sqrt{\frac{\sum_{i=1}^{n} d_i^2}{n}} = \sqrt{\frac{255}{40}} = 2.5$$

one-sixteenth of an inch, and are so recorded. The table shows the results of 40 observations. The mean of the observations was 72 units. The standard stimulus was 80, and thus the size of the illusion was 8 units (80 − 72 = 8).

Table 8.3 also indicates the calculation of the standard deviation, which can be used as an index of the discrimination threshold. The standard deviation = 2.5 units. Keep in mind that this will not equal the discrimination threshold in magnitude, but it is linearly related to the threshold and, if used consistently, will serve as a reliable index of the threshold.

Tracking methods of determining thresholds

Psychophysical methods often make direct use of only a small part of the observed data in deriving either the detection or the discrimination threshold. Notice, for example, that with the method of limits you only use the data points where shifts in judgment take place. The rest of the observations are just lost effort.

Tracking methods save work by taking readings only from stimuli near the threshold value. These methods are relatively new. As soon as the subject makes a judgment of "greater," the physical stimulus value is shifted downward, and upon a judgment of "less" it is shifted upward. Thus the measures are concentrated around the transition points, and the experimenter gets many more threshold estimates for a given amount of effort than with the older methods.

Key Ideas Box 8.3: Main methods for measuring discrimination thresholds

METHOD OF LIMITS

A *comparison stimulus* is varied in small, uniform steps sometimes from above and sometimes from below a *standard stimulus*. Both stimuli are presented on each trial. The subject has either two ("Greater" or "Less") or three ("Greater," "Less," or "Equal") response categories. The point on an ascending series where the subject first says "Greater" and the point on a descending series where the subject first says "Less" provide the basic data. The arithmetic mean of the transition points to "Less" judgments and the arithmetic mean of the transition points to "Greater" judgments will provide relatively stable estimates of these points of transition. The interval between these two transition points is one in which the comparison is neither clearly greater nor clearly less than the standard. It is called the *interval of uncertainty*. The midpoint of the interval of uncertainty provides an estimate of the point at which the comparison and the standard appear to be equal, the *point of subjective equality* (PSE), which is usually not the same as the point of true physical equality. The discrimination threshold is the difference between the PSE and the point where the comparison stimulus first appears to be different, either greater or less. In other words, the *discrimination threshold* is equal to one-half the interval of uncertainty.

METHOD OF ADJUSTMENT

The subject adjusts the comparison stimulus until it appears equal to the standard. This is done repeatedly, and the arithmetic mean of settings is the point of subjective equality. A statistical measure of the variability around these judgments of equality is taken as the discrimination threshold. For example, the *probable error* (0.6745 times the standard deviation) may be used.

METHOD OF CONSTANT STIMULI

A fixed set of comparison stimuli, presented in varying order, is compared to a standard in repeated trials. In one variant of the method, the subject has three response categories ("Greater," "Equal," and "Less"). A graph is made with the physical stimulus values on the horizontal axis and the proportion of judgments on the vertical axis. Separate plots of the proportions of "Greater" and of "Less" judgments are placed on the graph, and a well-fitting line is drawn through the plotted points, yielding two curves, as in Figure 8.5. The stimulus interval betweeen the point where the proportion of "Less" judgments equals 0.5 and the point where the proportion of "Greater" judgments equals 0.5 is taken to be the interval of uncertainty. As with the other methods, the midpoint of the interval of uncertainty is the point of subjective equality, and discrimination threshold is one-half the interval of uncertainty.

By custom, we call this a *tracking* method if the stimulus values vary continuously. If the stimulus values vary in discrete steps, the method is called the *up-and-down* or *staircase* method.

RELATIONSHIP OF THE JUST NOTICEABLE DIFFERENCE TO THE VALUE OF STANDARD STIMULUS: WEBER'S LAW

The size of the just noticeable difference (jnd) is not constant for all values of the standard stimulus. If a room is illuminated by one candle, added light from another candle is likely to be quite noticeable, but if there are 1000 candles in the room, an observer might well fail to notice the increase of light from one more. Many candles would be needed before an observer would notice a change in the illumination of a room already lighted by 1000 candles. This means that physical size of the jnd grows as the standard stimulus is increased. It is reasonable to ask the question whether there is any *systematic* relationship between the amount of increase in the standard stimulus and the corresponding amount of increase in the jnd.

An answer to this question was provided by one of the earliest of quantitative psychological laws—*Weber's Law.* This law states that jnd $= ks$; that is, the just discriminable amount of physical energy (jnd) is a constant proportion (k) of the standard (s) against which it is compared. This means, for example, that if an increase of 5 light units is just discriminable when the standard is 50 units, 10 will be a jnd when the standard is at 100. Then 100 will be the jnd when the standard is 1000, and so on. The law can be improved by adding a correction for the detection threshold. In this form it is $\text{jnd}/s = k + a$, where jnd = just noticeable difference, s = standard stimulus, k = a constant of proportionality, and a = the absolute threshold. The "law," especially in the latter form, is a good approximation of the actual state of affairs, at least over the middle range of stimulus values.

WEBER'S LAW: A WORKED EXAMPLE AND AN EXERCISE Weber's Law was a milestone in psychology because it permitted us to reach a great goal of science: to predict the value of measurements not yet observed. If you go back to the worked example of measuring the threshold for the discrimination of thickness, you can predict the difference thresholds that would occur with different thicknesses of the standard.

The discrimination threshold found in that experiment was 1.8 sheets. The standard stimulus was 7 sheets thick. Weber's Law says that

$$\frac{\text{jnd}}{\text{standard}}$$

is a constant. Thus,

$$\frac{1.8}{7} = k = 0.26$$

The number 0.26 is called Weber's fraction, and it should hold constant over wide variations in the standard stimulus. What would be the jnd be if you were to measure it with a standard equal to 10 sheets? The answer can be obtained by calculating:

$$\frac{x}{10} = 0.26;\ x = (10)\ (0.26) = 2.6$$

As an exercise, use the method of limits again to determine the difference threshold with the new standard stimulus. I went back and did the appropriate threshold measurements and got a threshold of exactly 2.6. Undoubtedly some luck entered into this precise confirmation of Weber's Law, since the method of measurement is not accurate enough to do that consistently. Nevertheless, in subsequent measures the results were always quite close to those predicted by Weber's Law.

If you do your own psychophysical experiments, you will probably be impressed that so simple an equation could predict what judgment you will make in the midst of all your doubts and confusions. Weber's Law, though hoary with age, remains an awesome and delightful accomplishment.

THE THEORY OF SIGNAL DETECTION

A major strategic change in psychophysics occurred fairly recently with the development of the theory of signal detection. This theory shifts attention away from the measurement of thresholds. Thresholds

Key Ideas Box 8.4: Weber's Law

A major goal of science is to predict measurements not yet made. Weber, very early in the development of experimental psychology, proposed a law to predict the value of unobserved jnd's after only one of them had been determined empirically. The law states that $\frac{\text{jnd}}{s} = k$. This means that the just noticeable difference divided by the standard stimulus is a constant. Thus, if one jnd has been established for a given dimension as 10 when the standard is 100, the value of the jnd for any other value of the standard can be predicted. For example, the jnd for a standard of 1000 should be 100 ($k = 1/10$). Weber's Law holds up rather well, especially for moderate values of the stimulus.

really are combined measures of two quite different things. One of them is *sensitivity* to stimuli or ability to discriminate stimuli. The other is the *criterion* according to which an observer decides whether a stimulus is present. In the measurement of thresholds, sensitivity and criterion are hopelessly intermixed. If the threshold for detecting a certain type of stimulus changed, we could not tell which of the two factors caused the change.

To look at a concrete example, let's take the case of changes in threshold for pain that may accompany acupuncture. We know that some people tolerate more intense noxious stimuli when they have been given acupuncture treatments. But is this because they actually feel less pained? If so, the change is due to a change in sensitivity. Or could it be that they are just less prone to report the pains they feel? They might simply be more relaxed and confident under acupuncture, with the same sensitivity. If so, the controlling factor is a change in criterion. (We discuss an actual study of this later.)

The theory of signal detection makes it possible to tell the difference between changes due to sensitivity and changes due to shifts in criterion. You can see how this might work by looking at Figure 8.8 and learning what it means. The figure shows a set of Receiver Operating Characteristic (ROC) curves. The probability of successfully detecting a stimulus when it is present, called a *hit,* is on the vertical axis. The probability of proclaiming a stimulus when none is really there, a *false alarm,* is on the horizontal axis. Do you notice the relationship between these probabilities and the type I and type II errors of statistical inference? Essentially, the problem of detecting a signal is of the same type as the problem of detecting real effects of an independent variable. The reasoning behind the theory of signal detection is borrowed from methods of statistical inference. But let's concentrate on Figure 8.8.

Look first at the diagonal line which splits the figure into two equal triangles. This line represents a case in which rate of hits is always equal to rate of false alarms. A person who was utterly lacking in sensitivity to the stimulus would have the two rates equal. For example, a completely blind person asked to detect small flashes of light might well be expected to respond in this way.

At the lower left corner of the diagonal line we start with hit and false-alarm rates both equal to zero. The blind person is saying "I don't see it" all the time; thus there are no hits or false alarms. But a blind person can get hits. Mere guessing will provide hits. As we move to the right on the diagonal line, hit rate and false-alarm rate go up, until finally they are both equal to 1.00—the person detects all signals but also has a maximal rate of false alarms. As we move from left to right on that diagonal, we see the effects of making the criterion increasingly lenient.

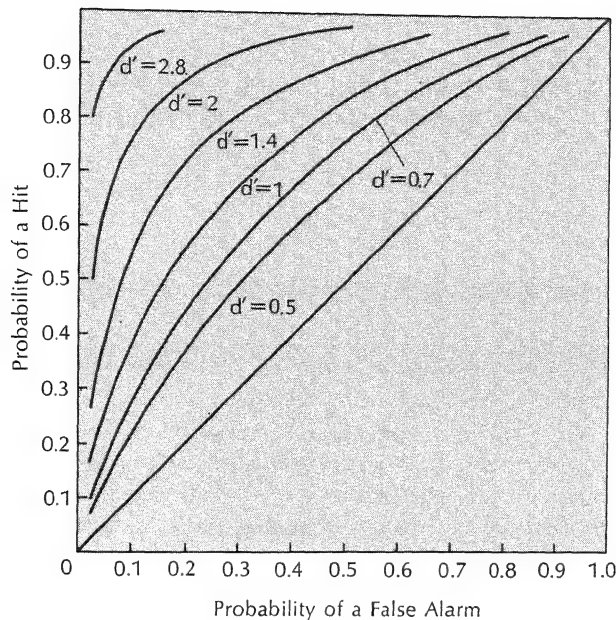



FIGURE 8.8 The receiver operating characteristic (ROC) curves. These curves show the probability of a hit plotted against the probability of a false alarm as the observer's criterion varies. The different curves are for various values of *d′*, a measure of the distance between the means of the signal and the signal-plus-noise distribution. The probability of a hit is the probability of accepting the hypothesis that a signal is present when it is correct that there is a signal in addition to the inevitable noise. The probability of a false alarm is the probability of accepting the hypothesis that there is a signal present when in fact there is only noise. (Modified from Swets, Tanner, & Birdsall, 1961.)

What will happen if there is a change in sensitivity? Then the observer can increase the hit rate without equally increasing the rate of false alarms. The curve will arch up toward the upper left corner of the graph. Notice that there are several such curves in Figure 8.8. Look at the very top curve, the one labeled "$d' = 2.8$." It starts with a probability of a hit equal to 0.8. What is the false-alarm rate at the same point? It is scarcely more than zero. This observer can detect stimuli and make few false alarms. The higher the curve in Figure 8.8, the greater the sensitivity it represents. We will discuss the meaning of "d'" later, but for now you can see that the larger it is, the greater is the sensitivity of the observer.

Even with greater sensitivity, there can be changes in criterion. Just as the more lenient criteria are represented by values to the right on the diagonal line, they are represented by values to the right on the curved lines above the diagonal. Notice that the curve with $d' = 2.8$ ends up with a hit rate of about 0.95. At the same time the rate of false alarms

goes up to about 0.15. Even a highly sensitive observer can improve the rate of hits by loosening up the criterion. But there is a price to pay in the increasing frequency of false alarms.

Let's take a look at the payoffs and costs involved in making various decisions with respect to signal detection. The following *payoff matrix* shows all the possibilities.

	Response of Observer	State of Reality	
		Signal	*No Signal*
Payoff Matrix:	Yes	Hit	False Alarm
	No	Miss	Correct Rejection

The observer has only two responses, either "Yes, there is a signal" or "No, there is no signal". A "yes" when there is a signal is a hit. A "yes" when there is no signal is a false alarm. A "no" when there is a signal is a miss. A "no" when there is no signal is a correct rejection. Each of these possible outcomes can be assigned a payoff or cost. What if I say to you, "I will pay you $100 for each correct hit and I will charge you 10¢ for each false alarm." What would you do? Most people would set their criterion at a very lenient level (you would be to the right on Figure 8.8). You can afford a few dimes for false alarms when you are collecting nice hundred-dollar bills for hits. Notice that the criterion depends on the payoffs and costs. If I turn it around and say you will get 10¢ for a hit and will lose $100 for each false alarm, your criterion will likely get very strict, and your data will be to the left of the graph.

A seductive approach to the theory of signal detection

A fanciful example applying signal-detection concepts to the maneuver of seduction may be useful in understanding the theory. The example is particularly useful in making the meaning of the ROC curve clear. So let us consider the interactions of Phil Faintheart and Bob Blunt with Vivian Vague and Felicia Frank.

Phil Faintheart and Bob Blunt both regard seducing a woman as a big payoff. On the other hand, Bob Blunt does not particularly mind being turned down, but Phil Faintheart regards a turndown as a mortifying catastrophe. Thus, the two men have very different criteria for deciding whether a signal is present or not. Phil sets his criterion very high, demanding a very clear signal before making an advance. Bob has a low criterion and acts on the slightest indication.

What will be the payoffs and costs of these two? Phil will rarely make a move unless a woman actually signals her receptiveness. In other words, his false-alarm rate will be low. However, his hit rate will also be low. He will miss signals when they are really there, and so his frequency of actually making a move when the woman is receptive will

be low. He will be a kind of living illustration of the adage: "Nothing ventured, nothing gained (or lost)." Both his gains and his losses will be small.

Bob Blunt, on the other hand, will make advances with only minimal stimuli. He will define as mere coyness the signals that Phil would take for a definite "no." When a woman is trying to tell him "yes," he will rarely miss it. His hit rate will be very high—but so will his false-alarm rate. He will commonly find himself in the predicament of having advanced upon a woman who had no interest in him.

Look at the ROC curves (Figure 8.8) now. The Phil Fainthearts of this world are represented to the left of the horizontal axis and the Bob Blunts are to the right. Actually the curves represent all gradations from the most cautious Faintheart to the grossest Blunt as they move from left to right. For simplicity, select the $d' = 0.5$ curve and ignore the others. By picking a point to the far left of the curve (say, where the probability of a false alarm = 0.1), you can see that the probability of a false alarm is low, and reading the hit probability from the corresponding point on the vertical axis of the graph, the probability of a hit can also be seen to be low (around 0.2). This is Phil Faintheart's score.

Let the place on the graph where the false-alarm probability is 0.7 represent Bob Blunt. The corresponding hit rate is around 0.8. Thus Bob Blunt "scores" a lot, but also makes many false alarms.

ENTER FELICIA FRANK AND VIVIAN VAGUE. Vivian Vague and Felicia Frank are represented by the different d' curves. Vivian either flirts with everybody without meaning it (plays Eric Berne's (1964) "rapo" game) or gives very few indications of interest in anyone. In either case, her signal-to-noise ratio is low. It is difficult to discriminate between her ongoing flirtation (noise) level and her real signals. The Vivian Vague class of "target" is represented by d′ curves toward the bottom of the graph of ROC curves (Figure 8.8). The $d' = 0.5$ curve we have been dealing with so far would nicely represent Vivian. A man will have to set his likelihood ratio very low if he is to approach her. His likelihood ratio is the amount of signal in comparison to noise that he requires before making a move. Notice also that Vivian can be vague either by flirting with everybody indiscriminately (even a strong signal might be missed if mixed with all that noise), or by being shy and quiet, giving signals so weak that they are hard to detect even when a prim woman is giving them. It is the ratio of signal to signal-plus-noise that makes Vivian Vague what she is, not the absolute magnitude of the signal.

At last we come to Felicia Frank. And we know exactly where she stands because she tells us in no uncertain terms. Even Phil has a hard time mustering a doubt about what Felicia has in mind. Felicia has a very high d' (let's say, $d' = 2.0$). Look back at the d' curve and notice

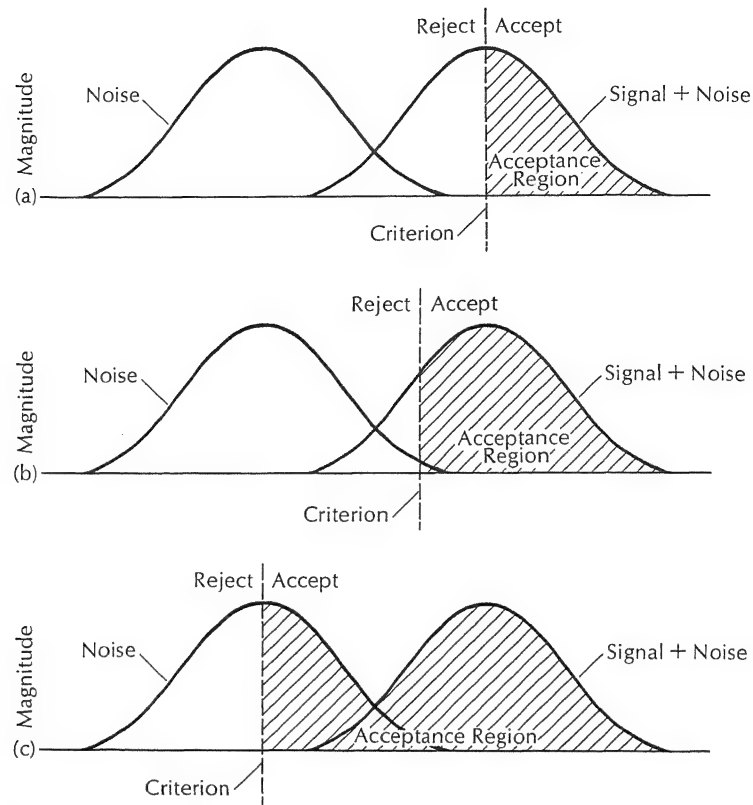



FIGURE 8.9 Effect of changes in criterion on rate of hits and of false alarms. Whenever some part of the signal + noise distribution is in the acceptance region, this indicates the occurrence of hits. Whenever some part of the noise distribution falls into the acceptance region, false alarms are indicated. (A) Stringent criterion: many signals will be missed, but there will be no false alarms. (B) Moderate criterion: fewer signals will be missed but some false alarms will occur. (C) Lax criterion: no signals will be missed but many false alarms will occur.

what happens to Phil when Felicia enters the scene. His false-alarm rates stays low. He is just as cautious as he ever was. But his hit rate jumps all the way to 0.8 or so. Felicia is just what Phil has been looking for. Bob Blunt's hit rate is not much improved by Felicia's openness.

A further impression of the meaning of the ROC curve can be gained from Figure 8.9. This figure shows how variations in the criterion lead to different hit and false-alarm rates. Two distributions are shown, one representing noise and the other representing signal plus noise. The detection process is assumed to be a statistical one and, though we have spoken simply of signal presence and absence, judgments are actually assumed to be made always in the presence of noise. The observer must distinguish between noise alone and signal plus noise.

It is assumed that this detection behavior is much like that of the experimenter making statistical inferences in order to decide whether an independent variable has an effect. The experimenter can get a large difference between the means of the experimental and the control group, yet fail to reject the null hypothesis. This will occur if there is a great deal of variability or "noise" in the experimental situation, especially if the criterion (alpha level) is stringent.

Thus, it makes little sense to rely on the magnitude of differences between means in making statistical inferences. It would make more sense to convert observations into some sort of "noise" unit. Just as we can convert feet to yards by dividing by 3, it is possible to convert raw measures to noise-unit measures by dividing by some unit of noise. In the theory of signal detection, the observer is assumed to base judgments on such a converted score. For a given observation, the score in question is the ratio of the value on the signal-plus-noise curve for that observation divided by the value on the noise curve for the same observation. This is called the *likelihood ratio.*

A given value of the likelihood ratio is adopted as an operational criterion for acceptance or rejection of the hypothesis that the signal is present. Under some conditions, the observer may require that the observation be many times the noise level before saying that the signal is present. Under other conditions, the observer may be more lenient. The choice of criterion will depend on such things as the values of hits and correct rejections and the costs of false alarms and misses. In any case, the criterion can vary, and as the criterion varies, the probabilities of getting a hit and of getting a false alarm will change. In Figure 8.9, the criterion is represented as a vertical line superimposed on the two distributions. If the likelihood ratio is anything to the left of the criterion line, the observer says "yes"; if it is to the right, "no." As the criterion line moves, the probability of saying "yes," given signal plus noise and the probability of "yes" given noise alone, will vary. It is the variations in the latter probabilities that are plotted on the ROC graph.

The probabilities in question will depend on the degree of separation between the noise and the signal-plus-noise distributions. For example, if there is a large portion of the signal–noise curve and a small portion of the noise curve over any given point on the horizontal axis, separating the two distributions may make the height of the noise distribution less for that point, and bringing them closer together may make it greater. Because the two "yes" probabilities vary with the degree of separation of the noise and signal–noise curves, there will be different ROC curves for the various separations. The graph of ROC curves (Figure 8.8) shows various curves for several separations. The separations are represented by a measure called d', which is simply the difference between the means of the noise and the signal–noise distribution expressed in terms of their standard deviation. It is d' that

is taken as the measure of sensitivity after the influence of the criterion has been taken into account.

Obtaining an empirical ROC curve

An ROC curve can be derived from theoretical considerations. On the other hand, such curves can be obtained empirically. Experimenters can use the curves to tell whether differences are due to variations in criterion or discriminability.

A given ROC curve results when the discriminability of the signal + noise distribution from the noise distribution stays constant and the criterion varies. An experimenter can vary the criterion by manipulating the payoff matrix. For example, one might give and take varying amounts of money for hits and false alarms.

For each criterion a hit rate and a false-alarm rate will be obtained, and these, when plotted on a graph, provide the points through which the desired smooth curve can be drawn. We know that a very large number of observations are needed to get stable results, so experimenters must spend a good deal of time and effort getting the curves by varying payoffs and costs.

Fortunately, there is a quicker alternative that is widely used as an equivalent method. The experimenter asks the subject to state on each trial a judgment from among a whole array of criteria. For example, the observer might be offered six categories: "Yes, positive," "Yes, fairly sure," "Yes, guess," "No, guess," "No, fairly sure," and "No, positive." A rough parallel of this *confidence rating technique* can be applied even in animal research by getting both an index of the animal's choices and an index of its confidence. The measures of confidence might be, for example, the latencies to responding on each trial.

Some applications of the theory of signal detection

The theory of signal detection is having a large impact on psychology. Psychophysics and its methods have always provided a model for the rest of psychology, and, with a certain time lag, the methods of signal detection analysis can be expected to pervade the rest of our field.

A major use of these methods derives from their capacity for distinguishing variations in sensitivity from variations in criterion. An ROC curve is an *isosensitivity* curve. If experimental observations when plotted fall on a given "same-sensitivity" curve, the results are presumably due to variations in criterion rather than to variations in sensitivity. If, on the other hand, points fall outside the "same-sensitivity" curve, the experimenter can reasonably presume that differences are due to variations in discriminability of the signal + noise from the noise distributions. Of course, error of measurement would make it unwise to attribute results to variations in discrimi-

nability versus criterion on the basis of single or even a few observations. In fact, statistical procedures provide the basis for judging whether there has been a significant variation in either d' or the measure of criterion from one treatment condition to another. The actual statistical procedures used vary with the type of problem under analysis. Descriptions of a number of them are to be found in Pastore and Scheirer (1974). But let's look at some real examples of signal detection theory as applied.

Clark and Yang (1974) attempted to learn whether acupuncture really lowered sensitivity to a defined pain stimulus (d') or whether it caused a change in criterion. They treated one arm of subjects with acupuncture and used the other arm as a control. Pain was induced by heating the arm (the radiant-heat method; see section on pain in this book). Comparisons were made before, during, and after acupuncture for each arm. They found that their acupuncture procedure was accompanied by a decrease in the proportion of withdrawals of the arm and of reports of pain. However, the change was due to modifications of the criterion and not to changes in sensitivity. They acknowledged that the generality of their finding should be tested across a variety of methods of acupuncture and types of painful stimulus. In contrast to the findings on the aftermath of acupuncture, analysis of the effects of nitrous oxide (Chapman, Murphy, & Butler, 1973) and aspirin (Dillon, 1972) revealed that these chemicals result in a decrease in sensitivity to stimuli that induce pain.

In a particularly interesting application of d' as a measure of perceptual sensitivity, Uhlela and Adams (1973) studied the relative ability of ethnic-minority and ethnic-majority children to discriminate situations that would likely be taken as justification for aggression. They found that d' was larger for minority groups. This indicates greater difficulty in discriminating "justified" from "unjustified" acts of aggression on the part of the children from minority groups.

They suggest two possible causes of this. Since "justification" was defined by the standards of the majority group, it might mean that barriers to communication cause the failures to agree on criteria of discrimination. A second possibility was that society as a whole actually pays less attention to the validity of justification when the aggressive acts are performed by children from minority groups. In any case, the results suggest the need for discrimination training (regarding both their acts and the persons who evaluate those acts) for children who get into trouble in, say, the school setting. Uhlela and Adams (1973) also found that children labeled "troublemakers" had larger d' values than other children. It would further be interesting to examine discriminations made by teachers. Do they change their criteria when dealing with differently perceived classes of children? Perhaps special training is needed for teachers and students both!

The memory operating characteristic (MOC) curve

One of the most important applications of the theory of signal detection has been in the area of memory. A memory "trace" can be viewed as a signal of a given strength. Then the basic reasoning of signal-detection theory is applicable. Curves similar to ROC curves can be drawn by plotting hit and false-alarm rates. The hit rate is based on the probability that a subject correctly identifies a familiar item. The false-alarm rate is based on the probability that the subject incorrectly identifies a new item as familiar. The MOC curve is also called an *isomnemonic* ("same memory") function. This is because it includes all the points possible with a single memory strength.

Empirical MOC curves can be constructed in basically the same ways as ROC curves. Because it is time-consuming to vary the criteria one at a time, experimenters usually prefer the confidence-rating technique. In such a case, subjects rate the identified material according to whether it is familiar or unfamiliar and also according to how confident they are that it falls into the chosen category of familiarity.

There are some problems in deciding on measures of sensitivity and criterion with MOC curves (and sometimes with ROC curves as well). These arise from doubts about the correctness of certain assumptions underlying the d' and β (criterion) measures. For example, an experimenter may not feel safe in assuming that the standard deviations of the noise and signal + noise distributions are equal. Yet this assumption underlies the use of d' and β . Measures that do not make these assumptions are available. They are discussed in Banks (1970).

There are two major applications of the MOC analysis. The first has been, of course, to provide measures of memory that are free of response bias. The second has been to tell whether given variations in recall are due to variations in strength of memory traces or differences in criterion. See Banks (1970) for a review of these and other uses of the MOC curve.

SCALING

The final problem of psychophysics is the problem of scaling. Scaling is also of general use in the practical matter of measuring psychological events.

Once you have a physical scale, you can do a variety of things with it. You can say that 10 pounds is twice as much as 5 pounds, that the distance from 5 pounds to 10 is equal to the distance from 10 pounds to 15, and so on. You can also write an equation showing how to translate values on one physical scale into values on certain other physical scales. For example, you can say that the length in inches can be obtained from the length in feet if you multiply the number of feet by 12. High-level measurement provides all these advantages. The physi-

Key Ideas Box 8.5: The theory of signal detection

Traditionally, the detection threshold was believed to be a fixed value. Variations in actual observation were thought to be due to error of measurement. Currently we recognize that the threshold varies as function of such things as payoffs for correct responses, costs of incorrect responses, and expectations due to frequency of the signal. The theory of signal detection replaces the concept of threshold with two concepts. These are *sensitivity* (usually called d') and *criterion* (usually symbolized as β). Variations in threshold may be due to either or both of these.

The *Receiver Operating Characteristic (ROC) curve* is a set of plots of probability of a correct detection of a signal (*hit*) as a function of the probability of saying there is a signal present when it is not (*false alarm*). If one has no sensitivity to the stimulus at all (for example, a blind person asked to detect tiny flashes of light), performance is represented by a diagonal originating at the intersection of horizontal and vertical axes (zero point; see Figure 8.8). At that point of intersection there are no hits and no false alarms. The person has no sensitivity and is not guessing. Further along the same diagonal line, guessing leads to hits, but also to proportional false alarms. These changes are entirely due to changed criterion (willingness to guess). Increases in sensitivity (d') lead to curves above the diagonal line. Rate of hits increases, but there is not a proportional increase in false alarms.

Signal detection provides a new psychophysical model that is having a great influence on psychological measurements. Various new applications of it to complex behavioral measurement outside of psychophysics are given in the text.

cal sciences advanced through the use of sophisticated measurement techniques, so it has been widely assumed that a science of psychology could probably profit from such techniques too.

Psychophysical and psychometric scales

There are fundamentally two types of scale in psychology. A scale is considered to be a *psychophysical scale* if it relates values on a psychological dimension to values on a physical dimension. Just as you can do a simple "transformation" (to go from feet to inches, multiply by 12), you should be able to go from physical to psychological scales by making a scale of each type of measure and determining the formula for transforming one to the other. Two great goals of

science are to measure events and to generate equations to predict things not yet observed. Hence the importance of psychological scaling.

Fechner's indirect method of psychophysical scaling

Gustav Theodor Fechner (1801–1887) made the first great attempt to create a psychophysical scale. He used an *indirect* method. We call his method "indirect" because he derived scale values from something else—namely, measures of discrimination. His fundamental notion was that the psychological distance between two things will be specified by the ease with which the two things are discriminated from each other. This seems reasonable. If two things are not discriminable, if observers regularly confuse them with each other, they must not be very far apart on a psychological scale. If, on the other hand, observers rarely confuse two things with each other, then the two things must be psychologically further apart. The degree to which two things are confused with each other (or, what amounts to the other side of the same coin, the extent to which they are discriminated from each other) is actually quite distinct from their distance apart on a psychological scale. This will become clear later. But Fechner believed that the location on such a scale could be *derived from* information about how frequently events were confused with each other. Thus his scale was indirect.

Fechner based his scale on Weber's earlier work on the just noticeable difference. Think of it this way: the point of subjective equality is the point where the measured events are most frequently confused. Thus, they must be close together. Moving outward from the point of subjective equality, we enter regions in which the measured events are less and less frequently confused. Since they are less frequently confused, they must be further apart on the psychological scale. With respect to the point of subjective equality and the difference threshold, Fechner would argue that we know from direct measurement that the comparison and standard stimuli are more often confused with each other at the point of subjective equality, but we know indirectly that they are also at that point less far apart on a psychological scale. When a politician says "There's not a dime's worth of difference" between the viewpoints of say, Democrats and Republicans, it seems to imply that, were he or she to look at a mixed list of statements and actions of members of the two parties it would not be possible to identify which statements belonged to members of which party. The statements and actions are readily confused with respect to source. It also seems reasonable to suppose that, if asked to place Democrats and Republicans on some scale based on his or her judgment (a psychological scale) of the distance between their political positions, the politician who made the statement would put them fairly close together. The discrim-

inability of two things does indeed seem to imply how far apart they will be if we are asked to estimate the distance as we experience it, psychologically.

Fechner extended this argument beyond the single jnd. If measures of discrimination show that two things are one jnd apart, this can be taken as a measure of their location on a psychological scale, and if they are two jnd's apart, they are farther apart than at one jnd. The jnd can in fact be used as a scale unit like the inch, the centimeter, or the gram. In order to place two events in proper position on a psychological scale, we could simply measure how many jnd's occur between them. We would then say something like: "Event A is 32.6 jnd's greater in intensity than Event B."

It would place an excessive burden on experimenters if they had to measure all the jnd's between two distant stimuli. Those of you who have measured the discrimination threshold know how much work it is. Imagine what it would be like to measure 300 or so of them just to place two stimuli on the scale! Fortunately, we have Weber's Law, and we need not make the actual measures. Weber's Law permits us to predict the jnd values and to save many an experimental hour.

Fechner wanted to get a psychophysical scale. He wanted to state an equation that indicated how the psychological scale related to physical scales. He started with Weber's Law, treated the jnd as a tiny unit on the psychological scale, and calculated mathematically what the desired equation must be. The resulting equation was as follows:

$$R = k \log S$$

The symbol on the left (R) is used to indicate the psychological scale value. The k is a constant of proportionality. It is just a conversion factor to make the scale unit sizes equal on both sides of the equation. For example, if you had "feet" on the right side of an equation and "inches" on the left side, you would have to put in a $k = 12$. The term "log S" indicates "logarithm of S," where S represents the physical scale value.

So the equation says that the psychological scale value equals a constant times the logarithm of the physical scale value. If such an equation works, we can predict how any physical event will be scaled psychologically, provided we know the constant of proportionality. One set of measures would have to be made in order to get the constant, but thereafter, all the future human judgments could be predicted mathematically. This is a very great scientific accomplishment.

For those readers who are not familiar with logarithms, the discussion of logarithms in Chapter 3 will help make clear the meaning of Fechner's equation (also called *Fechner's Law*). It is important to look over the tabulation in Chapter 3 carefully. Notice that the size of the logarithm grows slowly relative to the size of the original number.

Fechner's Law says that the psychological value grows as a logarithm while the physical value grows as the original number. It therefore indicates that a great deal of physical change must occur for a small amount of psychological change to occur. It implies, for example, that to double the loudness of your stereo, you have to increase the physical output 100-fold.

Fechner's Law is important. Like all great scientific laws, it allows us to predict events not yet observed. However, the events in question are human judgments, human experiences. Generations of humans believed these to be unpredictable. So Fechner's accomplishment was all the greater for having been believed beforehand to be impossible.

At a level less grand, Fechner's Law is of practical value. If we are to design stereos, lighting systems, acoustically sensible auditoriums, hearing aids, and the like, we need such equations to help us predict what will work prior to the heavy investment of resources.

And Fechner's Law *does* work to a nice degree of approximation. It stood unscathed for about 100 years before someone found a better law. And the better law does not say that Fechner's Law was not good; only that in 100 years we were finally able to improve on it.

Direct methods of psychophysical scaling

If Fechner's indirect method fails us, what alternatives do we have? A leading contemporary psychologist, S. S. Stevens (1961), has proposed and strongly defended the use of direct methods of scaling. Suppose you are persuaded by an experimenter to participate as a subject in an experiment that involves the scaling of pain. The experimenter places a little thing that looks like a miniature metallic scrub brush on your upper arm, then wraps it in a blood-pressure cuff. Inflating the blood-pressure cuff can vary the amount of pressure behind the little brush, and the apparatus can measure that pressure fairly exactly.

"In a minute, I will inflate the cuff," the experimenter says. "I want you to tell me the amount of pain you feel by giving it a convenient number. We call the number a modulus. Once you have picked it, I will present you with a series of pressures both above and below the modulus. They will be presented in irregular order. I want you to give each of them a number proportional to the pain you feel.

"Suppose you pick '10' for the modulus. Ask youself for the rest of the stimuli, 'If the modulus was 10, how much is this stimulus?' It's o.k. to use fractions or decimals.

"Oh, and by the way, don't worry because I will only give you about seven stimuli in any one session," the experimenter points out reassuringly. "We find that we get better data that way anyhow!"

That method of scaling is called the *method of magnitude estimation.* Notice that the experimenter is not figuring out the scale values

Key Ideas Box 8.6: Indirect methods of scaling and Fechner's Law

Scaling methods may be divided into two types—the indirect and the direct. Indirect methods are so-called because they derive scales from measures of something else. To illustrate, it is possible to scale physical things psychologically if we know how readily they are confused. The assumption is that two things that are frequently confused are perceptually closer together than two things that are seldom confused.

The discrimination threshold is a measure of the extent to which comparison stimuli are confused with (judged neither greater nor less than) a standard. Thus a psychophysical scale can be derived from discrimination thresholds.

Fechner derived his indirect psychophysical scale from a law that predicts discrimination thresholds—Weber's Law.

Fechner's law is stated:

$$R = k \log S$$

(Response equals a constant times the logarithm of the physical stimulus)

Fechner derived this equation by assuming that the discrimination threshold was a unit of sensation and that a just noticeable difference (jnd) would always give rise to the same sensation magnitude, no matter whether the standard was high or low on the scale. Fechner's Law is accurate to a good degree of approximation. It has been criticized on the grounds that (1) a just noticeable difference (or equal numbers of just noticeable differences) gives rise to different sensation magnitudes at different levels of the standard and (2) a different function (Stevens' power function, which is discussed later) predicts actual sensation magnitudes more accurately.

for the sensations from something else, but is asking you to give your preceived values directly.

Subjects have a great deal of freedom with the choice of numbers. This is true even when the experimenter assigns a number for the modulus. The results tend to be quite variable, and we usually gather data from 10 or so subjects before satisfying ourselves. The distribution of numbers tends to be skewed, so either the median or the

geometric mean—more often the geometric mean—represents the data best. (See J. C. Stevens et al., 1965, and Poulton, 1968, for practical details in using the method of magnitude estimation.)

There are a number of other direct methods of scaling. A couple of commonly used ones are the *method of magnitude production* and the *method of cross-modality matching.* With the method of magnitude production, the experimenter calls out numbers and you adjust the stimulus to match the numbers. It is obviously a great deal like the method of magnitude estimation. In our example, if the experimenter had chosen the method of magnitude production, you would have controlled the pressure in the cuff, and the experimenter would have said things like "If the modulus was 10, give me a stimulus equal to 15."

With cross-modality matching, the numbers are eliminated. Instead, you match the intensity of another stimulus to the one being scaled. For example, you might be given control over the intensity of a light and told to brighten or dim it till it matched your perceived intensity of pain.

Direct methods of scaling give results slightly at variance with Fechner's Law. Stevens maintains that the appropriate equation is not a logarithmic function, as Fechner believed, but a power function. We won't worry about the mathematical subtleties here; but whereas Fechner's equation was

$$R = k \log S$$

Stevens' equation is

$$R = kS^n$$

For both equations, R = the response of subject in psychological units, k = a constant of proportionality to correct for the sizes of psychological versus physical scale units, and S = the intensity of the physical stimulus in appropriate physical units.

The only symbol with which you are not familiar is n. To find n, we plot the psychological magnitudes against the physical magnitudes on graph paper with log-log coordinates. This simply means that the vertical and horizontal axes are each spaced out in units that vary logarithmically instead of arithmetically. It can be shown mathematically (see, for example, J. C. Stevens et al., 1965) that any data that follow the Stevens equation will form a straight line when plotted on log-log coordinates. Figure 8.10 shows several such plots.

But how does this give us n? It's simple. The value n is simply the slope of the resulting line (see Chapter 3 for a discussion of the slope of a line). Many different types of sensation have been investigated and had the exponent n determined for them. Table 8.4 shows the values of n for a number of them.

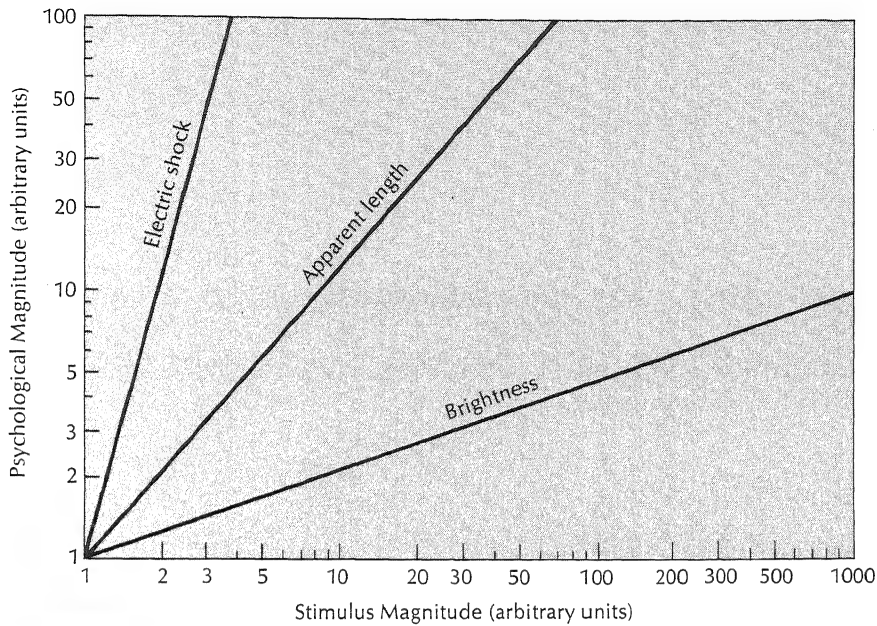



FIGURE 8.10 Scales of apparent magnitude for three continua plotted in log-log coordinates. (Reprinted from Stevens, in Rosenblith, 1961.)

TABLE 8.4 Representative Exponents of the Power Functions Relating Psychological Magnitude to Stimulus Magnitude on Prothetic Continua

Continuum	Exponent (*n*)	Stimulus Conditions
Loudness	0.6	Binaural
Loudness	0.54	Monaural
Brightness	0.33	5 deg target, dark-adapted eye
Brightness	0.5	Point source, dark-adapted eye
Lightness	1.2	Reflectance of gray papers
Smell	0.55	Coffee odor
Smell	0.6	Heptane
Taste	0.8	Saccharine
Taste	1.3	Sucrose
Taste	1.3	Salt
Temperature	1.0	Cold, on arm
Temperature	1.6	Warmth, on arm
Vibration	0.95	60 cups, on finger
Vibration	0.6	250 cups, on finger
Duration	1.1	White-noise stimulus
Repetition rate	1.0	Light, sound, touch, and shocks
Finger span	1.3	Thickness of wood blocks
Pressure on palm	1.1	Static force on skin
Heaviness	1.45	Lifted weights
Force of handgrip	1.7	Precision hand dynanometer
Autophonic level	1.1	Sound pressure or vocalization
Electric shock	3.5	60 Hz, through fingers

Key Ideas Box 8.7: Direct methods of psychophysical scaling and Stevens' Power Law

Fechner derived scale values indirectly from a measure of confusion, the jnd. Direct methods require the subject to assign numbers to scaled items directly. The *method of magnitude estimation* entails asking subjects to assign numbers to stimuli that reflect their magnitude on the dimension of interest. Any number, no matter how large or small, including fractions and decimals, can be used. Sometimes one item is assigned a value by the experimenter (this value is called a modulus), but not necessarily. With the *method of magnitude production*, the experimenter assigns numbers and the subject varies the stimulus to match the numbers. With *cross-modality matching* numbers are eliminated. One stimulus is varied to estimate the magnitude of another (for example, "Make this light as bright as the pain you feel").

Direct methods give stable results that fit *Stevens' Power Law* rather than Fechner's Law. Stevens' Power Law is $R = kS^n$, which means that the response is a constant times the stimulus raised to a power. The exponent n must be determined empirically for a given dimension (see Table 8.4).

Psychometric scaling

Often we want to scale a psychological dimension for which there is not a measured physical counterpart. For example, we might want to place the severity of different crimes on a scale (Coombs, 1967) or the relative quality of different composers of classical music (Fogelmann, 1933), or the vividness of memory images (Galton, 1883).

Clearly the most widely used procedure for psychometric scaling today is the method of *rating*. You have undoubtedly used rating scales many times in the past. You have probably also been rated on such scales. A very common scale is the *Likert-type rating scale*. With this type of scale, one is asked to indicate the degree to which one agrees or disagrees with various declarative items. For example, you might be given the statement: "The space program was a foolish waste of taxpayers' money." The categories of response might be

1. strongly agree
2. agree
3. undecided
4. disagree
5. strongly disagree

I have called this a Likert-*type* scale because a full-fledged Likert scale takes this form, but involves one more step. The experimenter must do an analysis of the items to be sure all items measure the same attitude (Kiesler et al., 1969). This can be done by determining which raters are in the top and bottom 25 percent for a given item with respect to their total score. Any item that fails to differentiate between people in the top and bottom 25 percent is then thrown out. In practice, experimenters often do not bother to do the analysis of items, but they use the Likert format. The resulting scales are satisfactory for most purposes.

Rating scales yield ordinal data. However, it is possible to derive interval data from them. We say that their *manifest content* is ordinal but their *latent content* is interval. The reasoning used is much like that of Fechner in that it derives the distance between ratings from the extent to which confusions occur. Just as Fechner used direct measures of the jnd to get indirect scales, we can use direct rating or direct ordinal measures to get, indirectly, interval data. The method for doing this can be found in Engen (1971).

THE METHOD OF PAIR COMPARISONS. A classic indirect method of interval scaling was the method of pair comparisons. It was honed and polished by Thurstone (1927). With this method, items are presented two at a time. Usually, all possible pairs are given. Suppose, for example, you wanted to know how people value various parts of the body. You might have "hands," "feet," "eyes," "tongue," and "ears" as stimuli. The various possible pairings would then be composed; for example, "hands-eyes," "feet-tongue," "eyes-ears."

At any given time, a subject would receive a single pair and be instructed to concentrate on that pair alone and decide which of the two was more valuable. This process would be repeated with the various pairs, and each pair would be judged many times. For close decisions there will be some proportion of judgments in which a given member of a pair is judged more valuable.

Essentially, Thurstone's argument was based on the notion that items are further apart on the scale if they are less easily confused. In fact, the *degree* of their confusion can be taken as a *measure* of their distance apart on the scale. Suppose, when choosing between "hands" and "tongue," that "hands" are rated more valuable 55 percent of the time. This implies that "tongue" is rated more valuable than "hands" 45 percent of the time. The decision must be a close one, and the two items must be close on the scale of value. In contrast to this, if one item were ranked more valuable than another on 90 percent of the choice trials, we would argue that the two items must be far apart. Note the similarity between Thurstone's reasoning and that of Fechner.

The principles behind Thurstone's reasoning are easily understood,

but considerable effort is needed to work out the details. Why would a person judge "hands" more important than "tongue" at one time and "tongue" more important than "hands" another time? Presumably a variety of chance factors contribute to these fluctuations. Any stimulus such as "hands" will produce a distribution of responses fluctuating around a given mean. This distribution Thurstone called the *discriminal dispersion* of the stimulus. What will be the form of the distribution? Will it be symmetrical around the mean or will it be skewed? It could take a variety of forms. Will all the various stimuli have equal dispersions (homogeneous variance)?

In order to get actual scale values, we have to answer such questions. Thurstone did so by making a variety of explicit assumptions. He stated his assumptions in the form of an equation called the *Law of Comparative Judgment.* We will not detail the law here. It actually took a variety of forms depending on whether various simplifying assumptions were made. Such details, and a clear treatment of the steps in getting actual scale values, may be found in Engen (1971).

The method of pair comparisons is a bit more cumbersome to use than many other methods of scaling. The number of possible pairs grows quite large as the number of items grows, so that it becomes impractical to work with more than 10 or 15 items. The calculations are also somewhat burdensome. Most researchers seem content to use simple rating scales and settle for measurement at an ordinal level.

RANKING METHODS. If judges are asked to place the items to be measured in order on the dimension of interest, it is called *ranking.* People sometimes confuse ranking with rating. To see the difference, suppose you want to evaluate the teaching of Professor John Smith. First you *rate* him on a Likert-type scale. Your alternatives are to rate him as

1. Outstanding
2. Above average
3. Average
4. Below average
5. Poor

Let's say you rate him 1, outstanding. Now suppose someone asks you to rank him along with ten other professors under whom you have taken courses. Even though you regard Smith as outstanding, you might find there are three other professors you feel are better. (Obviously you go to an exceptionally good school.) Thus Smith *ranks* number 4 in 10. Differences between rating and ranking can be marked.

Key Ideas Box 8.8: Psychometric scaling

Psychometric scaling involves measurement of psychological events that have no obvious physical counterpart (as they do in psychophysical scaling). Examples might be the seriousness of crimes, the beauty of music, or the vividness of memory images. We have no physical scale of these.

There are indirect and direct methods of psychometric scaling. Indirect methods assume that things that are easily confused with each other are closer together than things that are rarely confused. An example of an indirect method is the method of *pair comparisons,* in which all stimuli are paired, with one of the pair being chosen as above the other. An interval scale is derived from the frequencies of confusion.

Rating scales are probably the most widely used. The *Likert-type scale* requires the rater to rate the extent of agreement or disagreement with a given item (for example, 1 = strongly agree, 5 = strongly disagree). *Ranking methods* require that stimuli be placed in the order of their value. Note that you might fail to differentiate items by rating them, yet when asked to rank them you might be able to place one over the other.

The methods of *magnitude estimation* and other direct methods for getting interval and ratio scales can be used with psychometric as well as psychophysical scaling.

might otherwise avoid. On the other hand, it is necessary to keep the whole stimulus array present, at least mentally, in order to do ranking. Thus, it may not be a feasible method if there are large numbers of items to be ranked or if judges are not familiar with all of the items.

Ranking provides ordinal data directly. But it may also be used indirectly to get data on an interval scale. If a number of judges rank the items, the mean rankings of items that are close together on the scale will tend to be close. The mean rankings of items that are far apart will tend to be far apart. The degree to which judges "confuse" rankings of items can be taken as an index of closeness on an interval scale (see Engen, 1971).

DIRECT METHODS OF PSYCHOMETRIC SCALING. Just as we can ask subjects to make direct judgments of equal-interval or even ratio scales in psychophysics, we can do so when there is no corresponding physical scale. Essentially the same direct methods that are used in psychophysical scaling can be used for psychometric scaling. For

example, the method of magnitude estimation can be used to scale the beauty of paintings or the value of bodily parts just as readily as it can be used to scale lengths of lines or brightnesses of lights.

Study Questions

1. Give supporting evidence for the following statement: "Vision is determined by something more than the nature of the physical energies impinging on the eye."
2. What are the central problems of perception?
3. Give evidence for the statement that perceptual processes are more than merely informative.
4. What are the text's seven major strategies for the study of perception?
5. What are the main problems studied in psychophysics?
6. Why is psychophysics worth studying?
7. Is the detection threshold a point? Define it.
8. What is the method of limits?
9. What are ascending and descending series, and why are they employed?
10. What is the method of adjustment?
11. What is the method of constant stimuli? How is the threshold derived from it?
12. What is the discrimination problem?
13. What is:
    a. the interval of uncertainty?
    b. PSE?
    c. DL?
14. Explain Weber's Law.
15. What is Fechner's indirect method of scaling? Fechner's Law?
16. Explain direct methods of scaling.
17. What is Stevens' Power Law, and on what evidence is it based?
18. Contemporary psychophysicists have become disillusioned about the value of the "threshold." How and why?
19. What is the theory of signal detection?
20. Explain the ROC curve.
21. What is the likelihood ratio?
22. What is d'?
23. Distinguish between psychometric and psychophysical scaling.
24. Explain the following:
    a. Likert-type scale
    b. ranking
    c. pair comparisons
    d. psychometric magnitude estimation

Exercise

1. The Size of the Afterimage. It is easy to see afterimages and their relationship to the distance of the backdrop. Out of opaque cardboard, cut a figure, and tape it over a flashbulb unit. In a dim room, stand about five feet from the flasher and get someone to flash it while you look at it. Arrange for a fairly clear surface to gaze at afterwards, and look at the afterimage. Estimate its size in some familiar unit, such as inches. Vary your distance from the surface over a considerable range, estimating the size each time. How does the estimated size vary as a function of the distance?

9 PERCEPTION
Selected Research

Nativism versus empiricism in perception: A strategic question

Is our manner of perceiving inborn, or is it acquired through experience? Although this question seems straightforward, it is not easy to answer. Even its meaning is not entirely clear. The influences of experience and of our genetic makeup are intimately interwoven. It is hard to imagine an experimental treatment that would allow the two of them to vary independently. Many psychologists doubt that any measurement operations could be found to distinguish the innate from the acquired. If there are no conceivable operations separating innate and acquired traits, the distinction between them is operationally meaningless. Thus there is simply no scientific question to be answered.

It is easy to use this sort of argument to dismiss the issue of nativism versus empiricism; yet psychologists have deep convictions about the relative importance of these two classes of variables. The empiricist tradition is powerful among English-speaking psychologists. These psychologists have given relatively little attention to genetic or inborn structural factors as determinants of perception.

The implications of this attitude are profound, both from a practical point of view and from a purely scientific one. Our concepts of the very nature of humans and of the social and economic conditions that suit us have been determined by a view of ourselves that says that the "mind is a blank slate, the brain a random net." Ardrey (1961) has pointed out how the radical environmentalist position leads to the notion that we are about as well suited to one social and economic form as to another. But if much in us is inborn, then we might tend to search for situations that suit humans well.

From the purely scientific point of view, the question of nativism versus empiricism is highly important. If we ignore it and emphasize the environmental variables to the exclusion of all others, we are leaving out a major class of variables that determine human behavior.

Questions of this sort, that have profound long-range import, are *strategic questions. Nativism versus empiricism* is strategic because—though there may be no simple measurement operation corresponding to "native factors," and no simple and separate operation corresponding to "environmental factors"—the cumulative effect of many experiments is to give scientists a certain attitude toward a whole *class* of variables. Will we go on concentrating almost all of our attention on the influence of environmental factors on perception? Or will we begin to study and include in our interpretations of perceptual phenomena those factors inherent in the structure of organisms?

The nativism–empiricism issue is important not only to psychology. It is also critical to those who are interested in brain circuit theory. Just as psychology has given almost exclusive emphasis to the influence of past experience on behavior, brain circuit theory has been under the influence of random-net theories of brain functioning. The random-net view is that the brain is wired in no particular order at birth and that it becomes organized as a result of individual experiences. If the random-net theory is correct, it would appear to be out of the question for us ever to predict behavior by knowing something about the brain (Sutherland, 1963). If each brain is wired idiosyncratically, according to individual experience, it would be hard to see how we could ever get enough information about any given brain to make predictions about behavior. Hence, it is more than ever apparent that the nativism–empiricism issue has great strategic importance.

BEHAVIORIST VERSUS GESTALT TRADITIONS

Though behaviorists have stressed the importance of environment in the determination of behavior, gestalt psychologists tend to think of the principles of perceptual organization as innate. Max Wertheimer, who founded gestalt psychology, wrote a classic paper on the determinants of perceptual organization (Wertheimer, 1923). He placed great

what Brunswik (1956) has called *ecological sampling.* Ecological sampling refers to a procedure in which a representative sample of cues taken from the natural environment of an organism is measured and recorded. Thus we measure their natural occurrence; and we determine the actual relationship between the occurrence of a given stimulus at the sense organ and the corresponding physical objects. Wertheimer's objection would not apply here for in this case there would be two independent observations, one for the independent and one for the dependent variable. There is no danger of pseudo-explanation under such conditions.

TWO CLASSIC STUDIES

Gottschaldt's evaluation of the role of experience in perception

Gottschaldt (1926) investigated the influence of prior experience on the perception of embedded figures. He showed his subjects figures that were later to be embedded, varying the number of exposures to the stimulus. He then determined whether frequency of exposure had an influence on detection of the figure after it was embedded. Examples of configurations used by Gottschaldt are given in Figure 9.1. The figure to be embedded is called "*a* figure," and the more complex figure in which the *a* figure was to be identified is called the "*b* figure."

One group of subjects received 3 presentations of the *a* figure and a second group received 520 presentations of the *a* figure. Later, all subjects were presented with *b* figures and asked to describe what they saw. Subjects seldom noticed the *a* figure, and the two groups did not differ in the frequency of recognition.

The same groups of subjects were later given, respectively, 3 and 20 additional exposures to the *a* figures. They were also given instructions

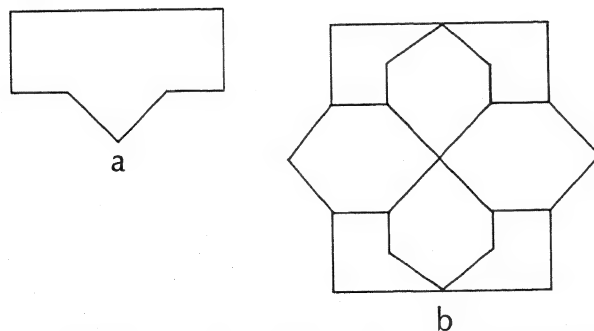



FIGURE 9.1 Examples of configurations (*a* and *b* figures) used by Gottschaldt (1926).

to look for the *a* figure in the *b* figure. The frequency of *a*-figure recognition increased with instructions, but the groups still did not differ in any noteworthy way.

This experiment has been regarded by some as the ultimate refutation of an empiricist view of perception. Yet the experiment is open to criticism on a number of very important points. A partial listing of these points follows.

1. It is designed to affirm the null hypothesis. What Gottschaldt has shown is that variation in the number of *a*-figure presentations does not have an effect on their subsequent identification. Experiments designed to affirm null statements are particularly error-prone, as was pointed out earlier, so it seems odd that Gottschaldt's finding should be allowed to occupy such a key position for nativists.

2. The assumption is made that 500 or more exposures would, if experience has a noteworthy influence on perception, cause a significant increase in the perception of the *a* figures. Whether or not this is true will depend on where the observer lies on the learning curve with respect to the figure. How often have adult subjects seen such figures? It is difficult to feel comfortable about the assumption that their learning curve has not already leveled off. If they have leveled off, then the empiricist position would lead us to expect that variations in frequency of *a*-figure presentation would be without influence. This is, of course, the same prediction made by nativism. Hence, the argument that Gottschaldt's findings permit us to decide between the two positions is shaky indeed.

3. No clear-cut reinforcers were provided for looking at the *a* figures; therefore we might interpret the *a*-figure presentations as *extinction* sessions.

4. In all fairness to Gottschaldt, it should be pointed out that *a*-figure perception appears not to have been at its limit. This is indicated by his own finding that ability to detect hidden *a* figures improved with practice and by similar findings of Hanawalt (1942), who also established that there was a kind of general improvement in the ability to detect such figures. However, this should be little consolation to the proponents of Gottschaldt, since the influence of learning on the perception of *a* figures is precisely what is being shown. Furthermore, it still remains an open question whether 500 exposures, especially unreinforced exposures, constitute a sufficient amount of practice to change figure perception in adults.

Gottschaldt's conclusions were challenged by the results of later experiments similar to his own. Djang (1937) had his subjects learn to associate a nonsense syllable with each stimulus. Both the *a* and the *b* figures were composed of dots, and the camouflage was generally less complete than in Gottschaldt's study. He found that the frequency with which the hidden test figure was discriminated varied significantly with prior practice.

Finally, Schwartz (1961) showed that prior experience had a significant effect on the perception of Gottschaldt's own figures when the subjects were trained to make color-name responses to the *a* figures. This should be enough to show that Gottschaldt's experiments hardly settled the issue of nativism versus empiricism. For further discussion of his experiments, see Postman (1963).

The visual response of previously blind humans

Whereas the nativistically oriented psychologists have given great weight to Gottschaldt's experiments, empiricists have leaned heavily on the work of von Senden (English translation, 1960). Von Senden did not conduct experiments at all. Rather, he gathered together all the material he could find on subjects whose eyes were restored after a period of blindness starting at or near birth. These individuals had generally not been completely blind, but rather had no *patterned* vision to speak of. This makes the case for empiricism a little more secure than if they had been completely blind, since it permits an empiricist account of at least minimal perception.

The recovered patients were able generally to tell figure from ground, but could not recognize such figures as circles or triangles except by procedures such as corner counting, and this only after prolonged practice. For the most part, they could not distinguish such grossly different figures as circle and triangle. Some patients could do this by painstakingly seeking corners. Even after long practice, the patients tended to fail to make what appear to be obvious generalizations from one setting to another, sometimes revealing that identifications were based on one or two gross cues. For example, having learned to identify a given object in white light, they might be at a loss to identify it in colored light, or having learned to identify it in one background, they might be at a loss to identify it in another. When permitted to use even the slightest auditory or tactile cues, they made accurate, swift identifications, indicating that the apparent visual deficits were not the result of some sort of gross emotional disruption.

Von Senden's data have been criticized on many grounds. Most important is that the reliability of the findings reported is not easy to assess. Many of the source materials are very old, and it is not easy to say whether observations were made in a manner that would satisfy our scientific standards.

Key Ideas Box 9.1: Classic attitudes toward and evaluations of the nativism–empiricism issue

Psychologists have long disagreed over whether perception is learned or innate. *Behaviorists* have tended to favor the notion that perception was based on learning; *Gestalt psychologists* came down on the side of the innate. Many psychologists have argued that the question is meaningless, on the grounds that the innate and acquired are so intermeshed that no measurement operations could distinguish them. Yet this "nature–nurture" issue is a *strategic question,* because attitudes toward it dictate whether whole classes of variables (for example, the genetic ones) will be studied.

Two classic studies had a great influence. On the nativist side, Gottschaldt varied the number of prior exposures to figures that he later embedded within other, more complex figures. Large variations in number of exposures had no influence on the frequency with which subjects detected the embedded figures. There are many objections to Gottschaldt's study, however. For example, he accepted the null hypothesis; he gave no reinforcers to maintain learning; he may not have given enough differential trials between exposed and nonexposed subjects. Furthermore, later investigators *have* shown an effect of prior exposures.

Von Senden compiled all the cases he could find of repair of the eyes in people who had been blind, or nearly so, since birth. He found that they had grave difficulty seeing even months after surgery. Each step of perception had to be acquired laboriously. This supported the empiricist view, since it seemed to indicate that perception is gradually acquired. However, these were poorly controlled clinical observations, and the perceptual retardation may have been due to damage to the retinas.

Furthermore, it would be wrong to suppose that a subject blind from birth has simply been isolated from the learning experiences on which perception might be based. There might, for example, have been neural degeneration due to disuse of the retina and optic pathways. Hebb (1949) has argued against this, maintaining that since neural regeneration does not occur in the central nervous systems of higher animals, and since the patients did eventually learn to perceive, the neural structures must have been healthy.

This argument seems odd because it has not been made clear whether the patients reached normal levels of visual functioning. It is

well known that recovery of function can occur after unregenerated damage to the nervous system through substitution of alternative neural systems. This could not occur if the retinas had been completely destroyed, but because of the residual vision present in von Senden's cases even before surgical correction of their condition, we would not expect that the retina could have degenerated altogether. Partial destruction could account for many of the postoperative effects.

Unfortunately, all these possibilities are mere speculations. For that reason, let us go on to more systematic researches that have been conducted to assess the influence of experience on perception. We have certainly not exhausted all possible objections to the von Senden studies, but since these studies are subject to the same objections that can be brought against "deprivation studies" in general, it is as well that a full discussion of the arguments be postponed until "deprivation experiments" have been described.

EXPERIMENTAL TACTICS USED TO EVALUATE THE ROLE OF EXPERIENCE IN PERCEPTION

Independent variables

An obvious way to determine whether perception is innate or learned is to vary prior experience in an experimental setting and to assess the influence of such variations on behavior. Indeed, this is the approach underlying the two classic studies just described (except for the fact that von Senden took advantage of what might be termed "ecological variation" in experience instead of manipulating it directly). Other experimenters have devised tactics that take advantage of the kinds of variations in experience that can be found in the natural habitat. For example, Mishkin and Forgays (1952) reasoned as follows. Assume that perception is learned. Since eye movements in reading English proceed from left to right, there should be a stronger tendency for words to the right of a fixated stimulus, rather than for words to the left, to be perceived. Such an effect actually occurs, but not until after about six years of schooling. Subjects whose first language is Yiddish recognize more Yiddish words to the left of the fixation point than to the right. The results seem to confirm the importance of learning, but only in a limited sphere of perception.

Most of the tactics we will discuss employ experimental manipulation. In some experiments, experience was varied *quantitatively.* For example, a number of investigators have reared animals in the dark from birth, and have subsequently submitted the animals to one or several behavioral tests that reflect perceptual ability. For the most part, a small amount of normal visual experience has been tolerated in order to allow an opportunity for care of the animals. However, in at

least one experiment (Nealy & Edwards, 1960) no patterned vision of any kind was allowed.

Experiments in which quantitative variation in experience is manipulated undoubtedly entail qualitative changes as well. We place them in a special category because only the quantitative variations are specified. There can be little doubt, however, that an animal reared in the dark has a special *kind* as well as a special *amount* of experience. For the most part, deprivation experiments have involved stimulus deprivation from birth, and this seems reasonable in light of the fact that the experimenter's intention is to argue that the performances of his subjects are due to innate factors. However, we can also introduce stimulus deprivation after an organism has a degree of perceptual experience, and it is interesting to compare the results to those that follow early deprivation. This has been done (Riesen, 1961) and will be discussed below.

Another approach to quantitative variation is that of measuring perception prior to experience by taking measures on newborn organisms. The limited behavioral repertoires of many newborns impose severe restrictions on this approach, but a good approximation to the ideal of measuring perceptual abilities at birth has been accomplished. (See, for example, Fantz, 1965.)

Instead of quantitative variation of perceptual experience, many experimenters have preferred to vary the quality of experience to which an organism is subjected, and then to look for differential effects on behavioral measures. In some of these experiments there may also be incidental quantitative variations. For example, animals have been reared in diffuse light without patterned visual experience. More narrowly qualitative have been the *rearrangement* or *disarrangement* experiments in which subjects, either from their earliest visual experience or later in life, are presented with distorted sensory inputs. The experiments of Held and his colleagues (Held, 1968), in which the presence of movement-produced changes in visual input is manipulated, also belong in the class of qualitative variation.

Dependent variables

Some of the most important tactical advances in our knowledge of the role of prior experience in perception have come from the discovery of new dependent variables. A major problem has been to distinguish between perceptual ability and such things as learning and motor abilities. Any measure of perceptual capacities that requires the organism to *learn* in order to indicate the ability to perceive necessarily confounds learning and perception. If perceptually deprived subjects learned a discrimination in an amount of time equal to that of normally reared subjects, then we could infer that no extra time was required for perceptual learning. However, any slowing of learning rates on the

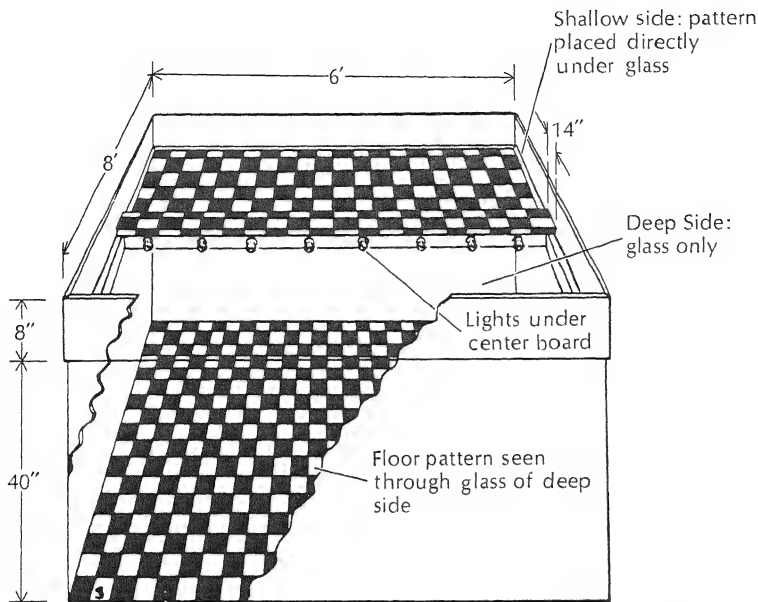



FIGURE 9.2 The visual cliff. See text for explanation. (From Walk, 1965.)

discrimination task would mean that the results were uninterpretable, since the outcome could be due either to impairment of learning ability or to an effect on perceptual processes, or both. For this reason, it has been of great value to use behaviors that are spontaneous, that occur without any special training. The *visual cliff* is a widely used apparatus that provides such a spontaneous behavior—the avoidance of what appears to be an abrupt precipice.

A picture of a visual cliff is given in Figure 9.2. It consists of a large box with a middle partition resembling a sort of bridge across the top of the box. It is safe for an organism to step off the bridge to either side of the partition because there is a glass surface covering both sides. However, one side "looks dangerous" because the floor is some distance below the glass. On the other side, the floor is flush against the glass; hence, that side appears shallow and safe. Subjects are placed on the middle bridgelike partition and a record is kept of whether they choose the deep or the shallow side when they step off the partition.

A very wide range of species has been studied on the cliff; in general, it can be said that they choose the shallow side as soon as they have the motor abilities necessary to make a choice. There may be exceptions, particularly those animals that have little reason to develop adaptive fears of precipices in their natural habitat (Gibson & Walk, 1960). Generally, though, no special training is necessary in order to get animals to choose the shallow side, and therefore side preference on

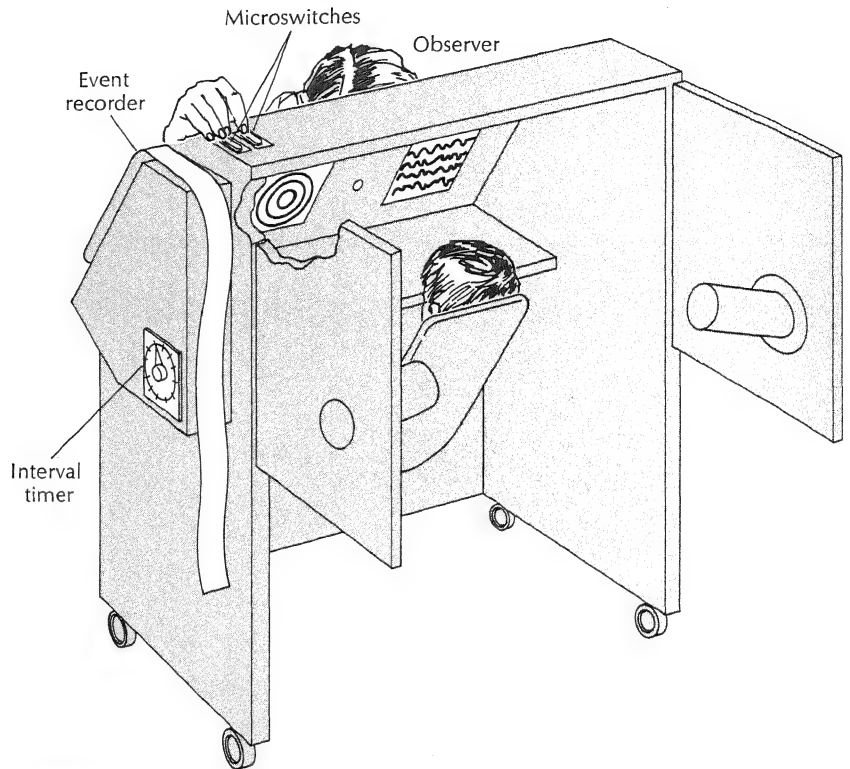



FIGURE 9.3 Drawing of infant visual-preference apparatus used for paired comparison tests of infants who were in semiupright position in a baby seat on the lap of a seated person. (From Fantz, 1967.)

the cliff is an excellent measure of visual discrimination with the learning factor absent or at least minimized.

On the other hand, a good deal of motor coordination and strength are needed to move about the cliff—more skill than many organisms have until some time after birth. We can minimize this motor component by finding a response that not only is spontaneous but also makes little demand on the motor system. A very important example is the *differential-fixation method* used by Fantz (1965), whose apparatus is pictured in Figure 9.3. The principle is very simple. The direction of the organism's gaze is used as a response, and it is measured by observing the images reflected from the eyes of the organism. It is possible to get adequate interobserver reliability with this technique; it is spontaneous and requires minimal motor skills.

One last major technique or set of techniques will be singled out before leaving the topic of tactics employed in evaluating the role of experience in perception. This is the physiological technique. Today

we have at least a partial understanding of the physiological and anatomical mechanisms that underlie the perception of visual shape. Just as, for example, the crystallographer has come to understand that there is a systematic relationship among patterns of shadows on his x-ray photographs and the scientific notions of molecules prevalent before crystallography, we have come to see a systematic relationship among the responses of single neural cells in the visual system and at least part of the process involved in the perception of shape (Hubel, 1963).

Though the knowledge gained about behavior through the use of physiological or anatomical techniques is imperfect, it can be of great use. It has, in fact, already served us well, as will be seen in the later discussion of physiological findings relevant to the nativism–empiricism issue. The main advantage of these techniques to date has been their ability to give us information about the structure of the shape-analyzing mechanisms of the brain without any use of the motor system. However, their potential applications are much greater than that, and have hardly begun to be realized. This is one of the many fascinating challenges still open to the experimental psychologist.

The list of techniques summarized here is far from complete, but it includes the most important modern techniques—those that in recent times have had a particularly powerful influence on our understanding of the role of prior experience in perception. Many other dependent-variable measures have been used, among which are, for example: the avoidance of obstacles, reactions such as eyeblink to approaching objects, visual placing (the subject is held in the experimenter's hands and is moved toward a ledge, which the subject anticipates by reaching out as it approaches), and the regulation of jumping force as a function of distance of visual objects toward which the jump must be made. We will go into a description of these and other techniques in the context of specified experiments of interest. The reader who is interested in a more detailed treatment is referred to Postman (1963) for a general treatment, and to Walk (1965) for a treatment directed specifically to problems involving visual depth and distance perception.

Quantitative variation through stimulus deprivation: Effects of dark-rearing

Studies in which animals are raised in the dark indicate that some species suffer marked visual impairment. Others suffer relatively little. In some cases (for example, in albino rabbits; see Walk, 1965), different investigators have come to opposite conclusions following dark-rearing of the same species.

Nealy and Edwards (1960) reared rats entirely in the dark, then tested them on a visual cliff. An important control group run by these investigators was a group of rats with eyes removed. Dark-reared rats

strongly preferred the shallow side, whereas blinded rats showed no preference.

In contrast to the rat experiments, there have been studies showing that some species suffer severe visual impairment following dark-rearing. However, the significance of these findings is diminished by several related observations:

1. Some animals showing visual impairment after dark-rearing have later shown improved performance despite the lack of any opportunity to learn the task. For example, Gibson and Walk (1960) reared kittens in the dark to the age of 27 days and found that the kittens at first showed no preference for the shallow side of the cliff, but began avoiding the deep side after a few days in the light. Instead of learning that the deep side was "safe," as a result of having stepped on it early in testing, they grew more and more shy of it. This seems to mean that the visual preference was inborn, but that dark-rearing disrupted performance temporarily.

2. Dark-rearing disrupts visual performance even in animals reared normally whose visual abilities were known earlier to be intact. For example, Riesen (1961) described a monkey that had been deprived of visual stimulation from ages 5–10 months and could no longer recognize food or food containers. In about 8 days the animal regained the discrimination. Riesen got similar results with a chimpanzee. Fantz (1965) applied his differential-fixation technique to the problem of measuring the effect of dark-rearing on visual preferences. He reported that visual interest in complex patterns *decreases* sharply during visual deprivation. This means that such deprivation impairs a capability that was at first present. It does not just preserve perceptual processes in their original state.

3. Dark-rearing has been shown (in at least one case) to result in physical degeneration of the retina. Riesen (1950) observed two chimpanzees reared in darkness and found that they failed to show any responses to complex patterns of light until after many hours in illuminated surroundings. However, their reflexes showed that they were sensitive to light. The retinas of these animals showed signs of deterioration.

Qualitative variation I: Effects of rearing in the presence of unpatterned stimulation

Since receptors deteriorate when external stimulation is virtually absent, why not measure the influence of prior experience by permitting stimulation while eliminating its patterned aspect? This has been done by a number of investigators. For example, Riesen (1950) continued his research by raising two chimpanzees with white contact

lenses over their eyes so that diffuse light stimulation was available but patterned stimulation was absent. Retinal degeneration did not occur under these conditions. Nevertheless, the animals showed deteriorated responding to visual stimuli.

Unfortunately, subsequent experiments indicated that even with the continuation of unpatterned stimulation, there are anatomical and physiological signs of deterioration in the visual system (see, for example, Wiesel & Hubel, 1965). Despite the apparent intactness of the retina, cellular degeneration and abnormal electrical responses occur in the central visual system of the brain.

Quantitative variation through measurement prior to experience

We have techniques for measuring behavior very early in the subject's life. We thereby minimize the chances for perceptual learning. This approach has a great advantage over either quantitative or qualitative deprivation experiments because it is not confounded by the possibility of damaging the organism physically.

The visual cliff has been used to evaluate perceptual abilities in species including chicks, turtles, rats, lambs, kids, pigs, kittens, dogs, and 6–14-month-old human infants (Walk & Gibson, 1961). All these species preferred the shallow side. In some cases this was measured on the first day of life and in others as soon as the organisms had the motor capacities necessary for testing.

There are obvious limitations to the conclusions we can draw from the cliff studies, particularly because of the developmental time lag necessary before more advanced species can be tested. Therefore, Fantz's differential-fixation method is important (see the section "Dependent Variables" and Figure 9.3).

Fantz has worked with monkeys, chimpanzees, and human infants. We will not go into the details of his work here; see his extensive review (Fantz, 1965) for such details. With the differential-fixation technique we can determine visual preferences in very young organisms, including human infants less than 24 hr old. Fantz measured visual preferences for stimuli differing along a number of cue dimensions (for example, shape, depth, texture), and he has even been able to make a good determination of visual acuity in human infants. Visual acuity can be measured because of the spontaneous preference of human infants during the first 6 months of life for a square of vertical black-and-white stripes over a gray square of equal size and reflectance. Visual preference should break down at the acuity threshold, where the stripes fuse into a blur. Human infants less than 24 hr old could see 1/8-in stripes at a distance of 9 in. The widespread view that the infant's world is a "blooming, buzzing confusion" cannot survive the onslaught of the data.

Qualitative variation II: Rearrangement and disarrangement

First let's try to answer the following question: Do organisms behave as though they were "prewired" when rearrangement or disarrangement of sensory input is given them before they experience a normal environment? Do organisms innately "know" what a normal environment should be like? With certain qualifications, the answer appears to be that they do. For example, Hess (1956) used prisms to displace vision 7 degrees to the side in newly hatched chicks. He found that pecking was accordingly displaced and accuracy did not improve even up to the point where the animals were near starvation.

Results obtained with kittens contrast to those obtained with chicks. Bishop (1959) raised kittens with vision contrived to be either upright or inverted. After the animals had learned a variety of mazes, obstacle courses, and an eye-paw coordination test display, subjects with inverted lenses were shifted to the upright condition, and vice versa. Readaptation was then followed by another series of tests. There were barely significant differences between the upright and the inverted conditions, and these were obtained only by combining the data from all behavioral measures. This suggests that in kittens (1) there is innate wiring; (2) it is very readily modified; and (3) its influence is slight when environmental pressures require a different perceptual organization.

Since the question of modifiability of prewired behaviors has come up several times in the previous paragraphs, let us face it directly: Are spontaneous behaviors modifiable? Notice that this question brings to light an important vagueness in the use of the word "innate." It takes little reading of the literature on nature versus nurture to discover that this word has a variety of meanings, often within the context of a single paper. Potentially, one might encounter behaviors that (1) occur without specific tutelage and cannot be changed through training; (2) are spontaneous, but which can be changed through training; (3) do not occur without specific training, but are irreversible; and (4) are learned and modifiable. Even this categorization may not do justice to the phenomena, since there might be gradations within categories. For example, there may be a continuum of differences in the *readiness* with which modifiable behaviors can be modified. We cannot draw clear conclusions about nativism versus empiricism as long as the concept of innateness remains so vague.

Now let us return to our clarified question: Are prewired behaviors modifiable? The Bishop (1959) study certainly indicates that they are, but other studies have led to the opposite conclusion. For example, Sperry (1951) has described a series of experiments involving rearrangement through surgical modification of a variety of organisms, followed by regeneration of sensory systems in abnormal patterns. Frogs, newts, and fishes with eyes rotated 180 degrees responded as

though the world were upside down. They persisted in doing so despite the maladaptiveness of the behaviors.

Pfister (1955) and Hess (1956) also reported failure to obtain adaptive modifications of the prewired behaviors of chicks; hence, the evidence suggests that there may be phylogenetic differences in this ability. However, more research needs to be done before we can draw this sort of conclusion with any confidence. If we say that readaptation cannot occur, we accept a null statement. Failure to get readaptation may simply be due to failure to use appropriate techniques. What appear to be differences between lower and higher species may in fact be differences due to variations in behavioral technique.

In most of the studies in which readaptation was not obtained, the conditions for retraining were far from ideal. For example, operant levels for the appropriate response are bound to be exceedingly low when subjects are put in a situation in which adaptive behaviors go strongly against their natural (or habitual) predilections. One would hardly be surprised to find that a rat failed to press a bar if it had previously learned to stay at the back of a Skinner box, away from the bar. It might well starve to death before coming into contact with the contingencies of reinforcement that the experimenter had arranged. The experimenter would hardly conclude from this that rats are incapable of pressing bars! Improved training techniques need to be introduced, presumably including a shaping procedure based on successive approximations to the correct behavior. At least one experiment indicates that chicks actually are capable of modifying their visual-motor coordinations when better behavioral techniques are introduced (Rossi, 1968).

REARRANGEMENT IN ADULT HUMANS. Another approach to the question of nativism versus empiricism involves the attempt to modify perceptual experience by distorting normal perception in adult humans by means of inverting lenses or displacing prisms.

Kohler (1962, 1964) and Smith and Smith (1962) have done extensive work with this method. Kohler kept the lenses on for very prolonged periods of time, sometimes several months. He also stressed the importance of motor activity, requiring his subjects to remain active and carry on normal activities. When inverting lenses were put on, the subjects became disoriented, with respect to both their phenomenal experience and their motor activities. Soon they made motor adjustments to the inverted world, performing such acts as fencing and bicycling fairly successfully. Eventually the world became phenomenally reinverted.

This reinversion took place in a piecemeal fashion. Subjects saw common objects as normally oriented while unusual objects were still seen as inverted. This piecemeal adaptation, on a smaller scale, was

commonly reported by earlier investigators. It is easy to understand if we view the perceptual reorientation as a process of developing new responses to various discriminative stimuli. Stimuli that are not often encountered do not have an opportunity to undergo the extinction and reconditioning of responses, which is necessary for reorientation to occur.

THE ROLE OF MOVEMENT IN ADAPTATION TO DISTORTION. There tends to be a relationship between the amount an organism moves while under perceptual distortion and the degree of readaptation to that distortion. Taking as a working hypothesis that organisms will readapt to parts of the environment present during movement, we can predict the outcomes of most known adaptation experiments. Can we say that if movement occurs, adaptation will occur? At least one modification of this position is necessary. The movement cannot be *passive.* It must be active. We are primarily indebted to Held and his associates for having established the importance of active movement in adaptation.

An excellent experiment to illustrate this point is the one done by Held and Hein (1963). In this experiment, kittens were raised in the dark until some degree of motor capability had been developed, and then pairs of the kittens were assigned respectively to an active movement and a passive movement condition. In both cases, the kittens moved while visual stimuli were present, but one kitten moved actively whereas the second kitten received an equal amount of passive movement; the amount of movement was equated through the use of a *yoked-control* procedure. A yoked control is one in which the behavior of one organism results in certain consequences for both itself and another organism that has no control over the consequences. In this case the active and passive cats were placed in what has been called a "kitten carousel," pictured in Figure 9.4. This consists of a turnstile with a fulcrum at the center, which permits horizontal circular movement of the cats. The cats are attached to the two ends of the turnstile, and the active one has its feet on the ground; the passive one has its feet on a false bottom so that its movements have no influence on the motion of the turnstile. Surrounding the turnstile and the cats is a cylinder covered with vertical stripes.

The kittens got all their visual experience in the carousel and, having reached maturity, received a number of behavioral tests. These included a tactile placing test, a visual placing test,[2] and a visual cliff test.

[2]In a "placing" test, the cat is moved toward a ledge while held by the experimenter. A normal cat will adjust its forepaws in anticipation of the ledge. This is "visual placing." If the cat fails at visual placing, it is moved forward until its forepaws touch the ledge. "Tactile placing" means that the cat moves its paws upward to place them on the ledge when they are touched by the ledge.

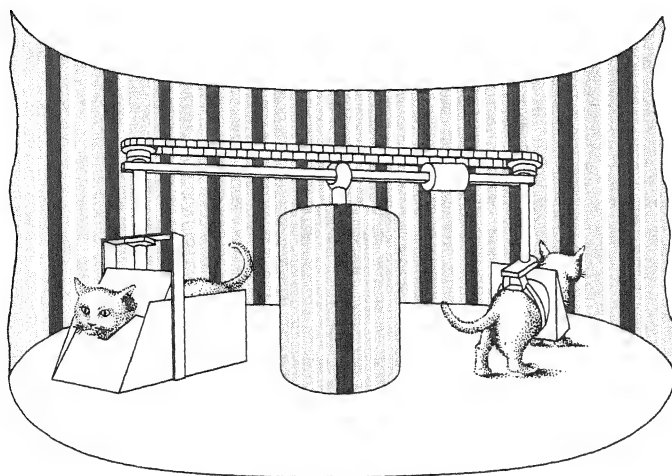



FIGURE 9.4 A kitten carousel. Movements of the "active" cat are transmitted to the "passive" cat. The passive cat receives as much input to the visual system as does the active one, but the passive cat's movements have no bearing on what it sees. (From Held, 1965.)

Active cats performed well on all tests. Passive cats, on the other hand, were able to perform well on the tactile placing test, but behaved virtually as though they were blind on the visually guided behavior tests. The fact that passive cats accomplished the tactile placing task indicates that the effects obtained on the other tests were not due to some generally disrupting factor such as impaired motivation or heightened emotionality.

Held and Hein (1963) concluded that active movement is necessary to the development of visually guided behaviors. However, it is important to note that the passive cats learned not to bother with responding to visual cues—in effect, they were on extinction. If there were a maturational process capable of producing visually guided behaviors, and if it were modifiable, the situation in which the visual world "moves away no matter what the organism does" might well reverse the normal process. It would be interesting to know whether animals already capable of visually guided behavior could lose it through such passive experience, and also whether visual recognition without gross movement (the sort that might be identified through respondent-conditioning procedures) remains in passive cats.

Held and Hein went on to use a within-subjects design for measuring the effects of active versus passive rearing. They used opposite eyes of the same cats. A monocular active and a monocular passive cat were paired for a time, and then the eyes and active-passive conditions were simultaneously switched. Essentially the same findings as in the previous experiment were found within animals. It is also of interest that microelectrode measures of the visual systems of these cats revealed no difference between active and passive eyes (Teuber, 1965).

CONCLUSION

Most psychologists have maintained that the infant's perceptual life begins as a "booming, buzzing confusion." That conviction was based largely on evidence from casual observation or from deprivation experiments. Casual observation is not sufficiently exact for scientific purposes, and deprivation experiments impair normal organismic functioning, even damaging the structure of the brain. We needed to develop new experimental techniques—to discover new independent and new dependent variables in order to make real headway in answering the questions that bear on the strategic issue of nativism versus empiricism.

The main methodological advances that have improved our understanding of the role of the prewired in perception are as follows:

1. The use of spontaneous behaviors, which make it possible to evaluate perceptual prowess without the confounding influence of learning
2. The development of perceptual measures, which both are spontaneous and make small demands on the motor skills of the organism tested
3. The growth in our knowledge of the physiological system underlying at least part of visual perception, which has made it possible to bypass behavioral measures altogether in evaluating the intactness of the system
4. The use of rearrangement techniques, which make it possible to assess whether perception is prewired, and the extent to which it is modifiable, without the use of deprivation as an independent variable

The development of new behavioral techniques often seems a humdrum task, but the impact of methodological discoveries can be immense. Think how long and how often William James's conception of the "booming, buzzing confusion" has been read, committed to memory, and taken as truth by students of behavior! If we had not developed new behavioral techniques, this misconception would likely have been perpetuated indefinitely. Consider also the important scientific "fallout" that comes as a side effect of looking at behavior in a new way, with new techniques. For example, physiological techniques have now made it clear that failure to exercise complex functions (in this case, pattern vision) can lead to the irrevocable impairment of the neural systems that underlie these functions. But this work can and should go far beyond the sensory analyzer parts of the brain. For example, a number of investigators in recent times have shown that there are fibers in the brain that respond specifically to

novel events (see, for example, Lettvin et al., 1961; Bettinger et al., 1967). Are these destroyed by rearing organisms in monotonous environments? If so, does this mean that the organisms lose their curiosity? This is not the place to make a list of the experiments suggested by developments in our knowledge of perceptual processes, but it should be clear that there is enough exciting work to keep generations of psychologists busy.

Before moving on to another topic, it is worth mentioning that one research area stands out as particularly in need of experimental attention. Most of the research we have discussed in this section has been on the innateness and the modifiability of visual processes. Admittedly, I have neglected a number of studies on other sensory systems, but the over-all impression conveyed here—that almost all work in the area is on vision—is correct. Maybe the study of other sensory systems will lead to quite different conclusions. Such study is almost certain to lead to important new insights. We need an evaluation of the generality of our present conclusions about the roles of the innate and the acquired.

It has been said that science progresses from the vague to the specific rather than (as has commonly been supposed) from the particular to the general. The nativism–empiricism question illustrates this point very well. In the beginning of our discussion there was much room for doubt about whether we had a worthwhile question at all. Having discussed a variety of research relevant to the nativism–empiricism issue, we find that there seems to be a noteworthy prewired facet of our behavior. However, there is not a great deal of solid evidence to support the view that perception is unmodifiable. Some investigators feel that the "visual analyzer" component of visual perception is fixed, whereas the relationship of vision to motion or the body's sense of position is plastic, but a greal deal more research needs to be done before this interesting hypothesis can be accepted wholeheartedly.

The conclusion drawn by Fantz (1965) puts our current state of knowledge in a nutshell: " . . . perception is innate in the neonate but largely learned in the adult!" Perception develops within the framework of prewired processes and its adult form seems heavily influenced by the framework in which it grows.

Pain

We have all felt pain, and to feel it is to gain a certain enthusiasm for finding out how to bring it under control. But pain has an inevitability about it. It is hard to believe that psychological procedures can bring it under control. Yet there is a great deal of evidence that pain can be

Key Ideas Box 9.2: Experimental tactics for deciding the nativism–empiricism issue

To decide on the issue of nativism versus empiricism, experimenters have modified the prior experiences of organisms or tried to measure perception early in life. The major modification of prior experience has been *stimulus deprivation*, especially *dark-rearing*. Many species, especially the more phylogenetically advanced, behave as one would expect if they had profound perceptual deficits after dark-rearing. Unfortunately, lack of stimulation is known to damage receptors, so the tactic is unsuccessful in resolving the issue. Rearing under *unpatterned stimulation* also can produce perceptual ineptness, but it has also been shown to produce damage to the brain. Perceptual deprivation is not a good tactic. Even if it did not produce damage, it might train the organisms to behave in maladap--tive ways that would give them the appearance of having perceptual deficits.

Attempts to measure perception early in life center on finding dependent measures that do not exceed the behavioral limitations of newborns, while revealing their perceptual abilities. The visual cliff (see Figure 9.2) requires a behavior that occurs without training in most organisms—the avoidance of an apparently deep dropoff. The technique of gaze fixation (see Figure 9.3) requires only movement of the eyes. Physiological measures with electrodes in the brain require no movements at all. All these methods give rise to results indicating considerable perceptual ability prior to opportunities for learning.

Studies of rearrangement and disarrangement involve modifying stimulus input, sometimes in a newborn, sometimes in an adult (for example, by placing distorting prisms over the eyes). Results, over-all, indicate great modifiability of perception. So current data suggest that perception is partially inborn, but readily modifiable.

controlled without the use of drugs or surgery. Psychological methods of controlling pain may one day largely supplant the ones in current use.

THE DOCTRINE OF SPECIFIC NERVE ENERGIES

One reason people believe that pain is inevitable is the view that pain occurs when specific "pain pathways" of nerves are active and that it does not occur when they are inactive. This view says that pain is determined by a single factor—the level of activity in pain pathways.

This is a prevalent view of the causes of all kinds of perceptual experience. It stems from a principle espoused by the great nineteenth-century physiologist Johannes Müller (1838)—the doctrine of specific nerve energies. This doctrine says that the quality of sensations is determined not by the quality of the stimuli but by the activity of the sensory systems that the stimuli activate.[3]

With the doctrine of specific nerve energies Müller improved upon the idea that a kind of pictorial representation of stimuli is transmitted to the mind. He emphasized the role of the interaction between physical and sensory systems in producing sensory experience. If the wrong stimulus reaches a sensory system, it will sometimes be perceived. But it will be perceived as having the quality that is characteristic of the sensory system, not of the physical event. Pressure on the eye can result in visual experiences, as anyone who has been given a hard punch in the eye can testify (you "see stars"). You can experience this less painfully by pressing gently on the side of the open eye. A kind of dark sunrise seems to occur in the visual field. Pressure is causing activity in the eye, where normally only light causes activity. The experience is not one of pressure, but one of vision. The sensory experience is characteristic of the sensory system, not that normally associated with the physical event. The brain does not passively carry pictures of the physical world into its higher recesses.

If the sensory system activated determines what will be experienced, we need only know about the sensory system's activity to have a complete account of the quality of experience. If this is true, vision will be due to activity in visual pathways, hearing to activity in auditory pathways, and pain to activity in pain pathways. Learning to control such experiences then becomes a matter of learning to control activity of the sensory system. In the area of pain, this means that attempts to control it will center on the development of techniques for blocking such activity, probably with drugs or surgery. And this is exactly where the traditional focus of attention has been.

MULTIPLE DETERMINATION

The doctrine of specific nerve energies tends to lead experimenters into research strategies that focus on single variables. When emphasis is placed on the multiple determination of experience, the research

[3]It could be argued that these are qualitatively different neural activities for each quality of sensation or, alternatively, that neural activity for different qualities of sensation differs not in quality but in spatial location (for example, the auditory nerve is spatially separated from the optic nerve). Müller was inclined toward the view that there are qualitatively different neural activities for different sensory qualities, but research soon revealed that neural activity is of the same quality regardless of its location. This issue is not critical to our discussion because it makes little difference in this context whether spatial or qualitative differences underlie the phenomenon.

strategy changes. A single sensory system does not, by itself, create sensory experience. Many different variables converge and give rise to experience. Thus to predict and control sensory experiences we must investigate a wide variety of independent variables. For example, motivation and attention are likely to have important effects on sensory experiences, including the experience of pain.

Some research has been done on the nonsensory determinants of pain, but the work has been relatively sparse compared to that within the traditional framework. I will describe some important research done with each of the two strategies.

PRODUCING PAIN IN THE LABORATORY

For the exact study of pain, scientists must induce it with measured stimuli. Many techniques have been used. Virtually any stimulus will produce pain if it is intense enough; bright light, loud sound, strong pressure, great heat, great cold, and so forth. However, certain types of stimuli have had wide use in laboratories. The following paragraphs contain a rough listing.

Mechanical stimuli

Von Frey (1895) used a mechanical method of inducing pain with what are known as "von Frey hairs." He attached horsehairs of various diameters and lengths to a lever, and the weights required to bend the hairs were determined on a balance. This gave him a means of measuring the physical stimulus needed to produce pain. More recently, Poser (1962) developed a more elaborate mechanical method. He used a standard blood-pressure cuff with a pressure gauge calibrated to 300 mm of mercury. With this device the pain stimulus was delivered by a flat acrylic base sewn into the cuff, which contains 94 pointed projections. When the cuff is in place, the projections rest against the surface of the subject's arm. The projections are sharp enough to cause most subjects to balk at a pressure of 250 mm, yet will not cause skin laceration until pressure exceeds 300 mm. Pressure can be quantified because air is pumped into the blood-pressure cuff in known quantities.

Another variant of the mechanical approach to producing pain involves inducing visceral (gut) pain. Chapman and Jones (1944) devised an apparatus consisting of a balloon that was inserted through the nose and secured in position 2–3 in. above the cardiac (lower) end of the esophagus. By expanding the balloon in a quantitative way, measures of the stimulus values producing pain could be made.

Thermal variations

Thermal variations have commonly been used as pain-producing stimuli. Goldscheider (1884) introduced heat as a means of evoking

pain. He transmitted the heat by contact, either through hot water or hot objects. In order to avoid the sensations of touch evoked by contact methods (temperature-inducing or mechanical), Alrutz (1897) used radiant heat by focusing the sun's rays on the skin. Hardy and his coworkers have used the radiant-heat method extensively (Hardy, Wolff, & Goodell, 1952). They focused the light and heat from a 500- or 1000-watt projection lamp precisely on 3.5 cm^2 of artificially blackened skin, usually on the forehead of the subject, for 3 sec. The current delivered to the filament of the lamp was varied.

The ultrasonic method

It is possible to produce deep pain with ultrasonic vibration. Ultrasound is a very deep heating procedure. Gelfand (1963) introduced ultrasound of fixed intensity and measured duration, thereby producing quantified pain stimuli.

Cold-pressor method

Cold can also cause pain. Hilgard (1969) has used this *cold-pressor* method in his studies of hypnotic modification of pain. The subject's forearm was placed in circulating water at low temperatures (for example, 0°C, 5°C, and so on) for about 40 sec, and various measures of pain response were taken. For example, subjects might estimate pain on a scale from 0 to 10, 0 meaning no pain and 10 meaning a pain so great that they wish to remove their hands.

Ischemic pain

Hilgard (1969) has also used an ischemic method of producing pain. Ischemia is deficiency of blood supply in a muscle or other organ due to constriction or obstruction of a blood vessel. Hilgard's method was to place a tourniquet just above the elbow and then ask the subject to squeeze a hand-grip device a standard number of times. After squeezing, pain begins to mount in the deprived muscles and can be measured in ways similar to those used with the cold pressor.

Electrical stimulation

Electrical stimulation of skin, teeth or nerves has been used to produce pain. Skin stimulation has a number of disadvantages:

1. Changes in skin resistance to the flow of current cause modifications in the patterns of current flow
2. Other sensory systems are heavily stimulated
3. External factors such as temperature and humidity are difficult to control
4. Internal factors such as temperature and circulation may influence measures (Mueller, Loeffel, & Mead, 1953; Goetzl, 1946)

Stimulation of teeth by delivering current through fillings has been used fairly widely. Uniform results can be obtained in the tooth electrode when placed at the same point on a carefully dried tooth (Bjorn, 1946). Stimulation of peripheral nerves has yielded valuable information (see, for example, Collins, Nulsen, & Shealy, 1966). It is, of course, not easy to get participants for such experiments!

MEASURING PAIN

The most common methods of measuring pain are used on contemporaneous verbal scaling of the intensity of the pain. Any of the familiar scales might be used. For example, a person might be asked to assign the intensity of the pain a number from 1 to 5, where 1 = mild pain and 5 = unbearable pain. Or, having been given a drug designed to relieve pain, the person might be asked to assign a fraction indicating whether the pain after the drug is 1/4, 1/3, 1/2, or the like, of the original pain.

Hilgard (1973) has used a rather unusual variant of the verbal report. He got subjects to report on their pain by way of automatic writing. Automatic writing is done by letting the hand scribble without giving any deliberate direction. It seems that some people engage in it without special training, but it can also be developed in most of us through a sort of shaping procedure (Mühl, 1972). Automatic writing seems to tap into aspects of a person's experience that are set off from ordinary consciousness (Mühl, 1972). Hilgard found that contemporaneous reports sometimes gave differing results depending on whether normal verbal methods or automatic writing were used. For example, a subject under hypnosis might indicate virtually no pain when asked to rate it in the usual way, but indicate severe pain by way of automatic writing.

We can also use physiological or behavioral measures of pain. Blood pressure typically goes up with increases in pain (though it drops at very high intensities, when the person begins to faint). Decreases of GSR and increases in muscular tension also can be used as measures. In fact, a variety of physiological measures, yet to be explored, hold promise as measures of pain.

Few investigators have thought to use nonverbal behavioral measures of pain. Sullivan (1973) devised such a measure. He induced ischemic pain in the arm. Having wrapped the arm in such a way as to squeeze the blood out of it and keep it out, he asked the subjects to lift a weight with the blood-deprived arm. This created an oxygen debt, which causes ischemic pain. When he took measures of pain (the experiment involved repeated measures on the same subjects) he asked for verbal ratings, took a variety of physiological measures, and asked them to lift the weight as many times as they would. The number of lifts provided a nice nonverbal, behavioral supplement to the other measures. More work needs to be done exploring the use of behavioral measures in research on pain.

Key Ideas Box 9.3: Measuring pain

Various methods have been devised for measuring pain in the laboratory. *Independent variables* include manipulation of *mechanical pressure* (for example, by putting an abrasive item beneath a blood pressure cuff and inflating it) and *thermal variations* (such as focusing a *infra-red heat lamp* on a blackened spot of skin or, for deep pain, using *ultrasonic heating*). A widely used thermal variation is the *cold pressor.* Part of the body is placed in a cold bath of specified temperature. The *ischemic* method involves cutting off blood supply to part of the body, producing oxygen deprivation. *Electrical stimulation,* though erratic, has sometimes been used effectively.

Dependent measures include the familiar *verbal scaling* methods (and some less familiar ones, such as automatic writing), *nonverbal behavioral measures,* such as the willingness to cause more of the pain-producing stimulus, and *physiological measures* such as GSR and blood pressure. *Multiple measures* are to be preferred.

When multiple measures of pain are used, they often do not correlate with each other as highly as we would like. If the term "pain" corresponded to a simple phenomenon, we might expect virtually perfect correlations between measures. But "pain" is a term with various meanings in ordinary language.

Melzack (1973) has pointed out the need for a more detailed breakdown of verbal measures of pain. We usually ask for measures of intensity. But pain varies along a variety of dimensions. We show this whenever we describe a pain as "pulsing," "dull," "shooting," "burning," "aching," "sickening," or "terrifying." As Melzack (1973) says, ". . . to describe pain solely in terms of intensity is like specifying the visual world only in terms of light flux without regard to pattern, colour, texture, and the many other dimensions of visual experience."

Melzack and Torgerson (1971) analyzed descriptions of pain into a variety of categories. These included sensory descriptions (such things as temporal, thermal, and spatial properties), emotional descriptions (pain may produce tension, fear, sickness, and so on) and evaluative descriptions (for example, mild, distressing, excruciating). The development of new dependent measures often stimulates many new discoveries, because whole classes of previously unmeasurable phenomena begin to manifest themselves. The work of Melzack and Torgerson (1971) and of Sullivan (1973) may be very useful to future researchers.

ANATOMICAL STRUCTURES UNDERLYING PAIN

Receptors

Various parts of the body surface produce pain differentially. You can observe this on your own body by pinching the back of the hand or the skin over the knee and comparing the resulting pain with that produced by pinching the back surface of the upper arm, over the triceps muscle. The area over the triceps is much more sensitive than other areas. With more refined techniques it is possible to show large variations in pain sensitivity within rather small areas of skin.

Pain is not alone in varying over the surface of the body. Other skin sensations vary in a similar way, though different areas respond maximally to different sensations. One could make a "cold map," a "pain map," or a "touch map" of the body's surface, and each map would be different. By inference to structure one might reasonably suppose that there are different types of receptor corresponding to the various regions of skin sensitivity. This view would also accord with the strategy stemming from the law of specific nerve energies, since the qualities of sensations should depend on the structure stimulated. Von Frey (1895) maintained that there were four distinct skin sensations (pain, pressure, cold, and warmth) and that each of these corresponded to an identifiable receptor in the skin. Since pain was the most widely distributed response, the most widely distributed "receptors"—free nerve endings that occur between epithelial cells—were assigned the role of receptors for pain. Krause bulbs just beneath the surface of the skin were taken to be receptors for cold; Ruffini endings, deep down, were for warmth; and free endings around the hair follicles were for pressure. Von Frey's view was in accord with the doctrine of specific nerve energies, and met a very receptive audience. His views were accepted rather uncritically, and even today many textbooks repeat them. Yet evidence against his view was available even before it had been published. Goldscheider (1884) had excised cold and warmth-sensitive spots, but when he studied them microscopically he found only free nerve endings (see Boring, 1942, for an account of this). In general, contemporary evidence has supported Goldscheider rather than von Frey. For example, Lele and Weddell (1956, 1959) have found the cornea capable of responses to touch, cold, warmth, and pain despite the fact that the cornea contains only free nerve endings. If von Frey were correct there should be various types of receptor in the cornea to mediate the variety of sensations arising there.

Peripheral nerve fibers

There is a great deal of evidence to show that there are two distinct sorts of cutaneous pain, sometimes referred to as *fast pain* and *slow pain.* Fast pain has, in addition to a short latency, a pricking quality. Slow pain is of a burning and less bearable quality.

Nerve fibers are classified, according to such factors as conduction velocity, into A, B, and C fibers, with A being the fastest and C the slowest category. These classes of fiber also differ with respect to such things as size (A fibers are of large diameter), the presence of a fatty "myelin sheath" around their surface, and their response to drugs. The A category includes a considerable range of velocities. It has been divided into a set of subcategories, labeled by adding letters of the Greek alphabet to the A. Thus there are large, fast, well-myelinated A-alpha fibers, A-beta fibers, and so on. Small myelinated A-delta fibers seem to play a special role in fast pain. Slow-conducting unmyelinated C fibers have a special relationship to slow pain.

Nerves are bundles of fibers. When the nerve bundle is placed under pressure, fibers are damaged and their conduction blocked. Larger fibers tend to be damaged first. This is because the small ones are better able to squeeze into the available space between large ones. When conduction in peripheral nerves is blocked by pressure, the double nature of pain persists as long as all fibers of A-delta size and smaller continue to conduct. When conduction in these small myelinated fibers fails, fast pain disappears, but slow pain persists.

Collins, Nulsen, and Shealy (1966) have done an especially straightforward study of the role of peripheral fibers in pain. They exposed peripheral nerves of human subjects under a local anesthetic and got reports of the sensory experiences resulting from direct nerve stimulation. By varying the intensity of the stimulating electric current, they were able to activate the large fibers first, then the smaller ones, and so on. When only the fibers larger than the delta ones were activated, no experience of pain occurred. When a stronger stimulus triggered the delta fibers, pricking pain was reported at frequencies of stimulation as low as 3 Hz. Next, conduction in myelinated fibers was blocked. A much stronger stimulus was used to activate C fibers. No sensation at all occurred when a single nerve-impulse volley was induced. But at frequencies of 3 volleys per second the sensation of pain was evoked. The experience was so severe that subjects refused to allow it to be done to them again.

Central pain pathways

We know the pathways and "centers" related to pain. They are too complex for detailed treatment here. Such information can be obtained in Mountcastle (1968). Briefly, severing the anterolateral spinal cord results in absence of sensation (including pain), beginning a few segments below the level of the cut. Stimulation of the anterolateral system results in pain in conscious humans. The projections go by way of two systems. The first route is through the neospinothalamic system into the ventrobasal and posterolateral thalamus and the somatosensory cortex. The thalamus is a sensory relay structure through which most sensory systems pass on their way to the cortex, the surface layer

of brain cells. An excruciating pain, which is particularly unmanageable, occurs when a branch of the posterior cerebral artery, the thalamogeniculate artery, becomes clotted, causing damage to posterior thalamic nuclei. This is known as thalamic pain. Many believe that no pain results from stimulation of the cortex. We owe much of our knowledge of the functions of cortex to the fact that neurosurgeons are able to stimulate it in waking humans and get their responses without particular discomfort to the subjects. Since the cortex is phylogenetically recent and pain is phylogenetically ancient, this could be taken to mean that pain is mediated only by lower, older brain areas. In reality, however, pain does occur upon cortical stimulation, though only in a small percentage of cases. The low probability of getting pain after cortical stimulation may be due to the fact that pain-inducing areas are buried in the infoldings of the postcentral gyrus (Mountcastle, 1968).

A second route for pain is by way of fibers that ascend near the middle of the spinal cord and end in the reticular formation and medial intralaminar thalamus and limbic system. Electrical stimulation of the tooth at painful intensities evokes activity in both the first and second projection systems, and damage to them may strikingly diminish pain perception and response (Melzack & Casey, 1968).

PARADOXES OF PAIN

The study of pain has paralleled rather closely the study of other perceptual experiences. Subjects give indications of their responses to physical stimuli. These stimuli are categorized and, by inference to structure, anatomical systems are hypothesized to underlie the responses. Just as studies of color mixing gave rise to the postulation of three pigments in the eye, the mapping of regional variations in sensitivity of skin gave rise to von Frey's (1895) conception of cutaneous receptors. Just as visual pathways were identified in the nervous system, pain pathways have been observed. The most beneficial work has inevitably been the direct physiological observation of hypothesized structures. Direct measures of color pigments and neural cells responding to various visual features have their parallels in the area of pain, and in both cases go a long way toward strengthening our belief that there is some truth in the inferences about structure that have been made on the basis of behavior.

Many variables can influence percepts. Signal-detection theory places emphasis on such factors as signal probability, payoffs, and costs. The threshold is not simply a function of receptor sensitivites. What parallels to these new viewpoints exist in the area of pain? If we maintain that the cause of pain is activity in certain sensory pathways, we are left with a number of paradoxes. Pain sometimes fails to occur when we have every reason to suppose that A-delta and C fibers are

active. For example, Beecher (1959) reported that most American soldiers wounded at the Anzio beachhead "entirely denied pain from their extensive wounds or had so little that they did not want any medication to relieve it." He thought this was related to the great positive consequence of injury—that of being allowed to get out of the battle. Wounds of comparable magnitude acquired in surgery are very painful. Weinstein (1968, p. 441) tells of a personal experience in which he was doused with gasoline and set on fire. In his own words,

> During the entire period of time, which occupied almost half a minute, I was completely aware visually that the flames were engulfing my body, but I had not the slightest painful or tactile sensation. I was able to put out the flames and did not experience pain at all at the time. I walked a quarter of a mile to the hospital, and during this period I never had any suggestion of pain. Subsequently, I had rather severe pain, was given morphine, and was hospitalized for about a month.

There can be little doubt that there was massive inflow in the A-delta and C fiber systems, but there was no pain.

Pain sometimes occurs when we have every reason to suppose that there is no more than normal baseline activity in the "pain fibers." Phantom-limb pain is a good example of this. It is not uncommon for patients to report pain in limbs that have been removed. Sometimes the pain is very severe. Phantom-limb pain is distinct from stump pain. It is well known that pain can be produced by physical characteristics of a stump and that this sort of pain is relieved by surgical modification of the stump. But the phantom-limb pain remains. In fact, "a review of the literature on painful phantoms soon shows that neurosurgical procedures are distressingly ineffective. Operations ranging from peripheral denervation to prefrontal leucotomy usually fail to produce the desired relief" (Sternbach, 1968, p. 130). On the other hand, psychotherapy is often successful in relieving the phantom pain (Sternbach, 1968). Melzack (1973) has written an interesting account of various types of pain that seem poorly related to the simple physical stimulus that presumably induced them.

PAIN AND HYPNOSIS

One of the earliest anesthetics was hypnosis (Boring, 1950), yet surprisingly little research has been done on hypnosis as an independent variable controlling pain. One problem with hypnosis as an independent variable is that it is difficult to produce reliably. There are large inter- and intrasubject variations in susceptibility to hypnotic trance. If we take the view that our independent variable is simply the procedure we use to induce hypnosis, without regard to whether the

trance occurred, we will have a reproducible independent variable. However, we will also have to face a good deal of variability in our dependent measures. If, on the other hand, we use, as most people have done, the trance itself as the independent variable, we will have to select our subjects nonrandomly. We will also have to face the problem of determining when a trance has or has not occurred, and whether what we call a trance is the same thing as what someone else calls a trance.

Despite the experimental difficulties, there has been some interesting work on hypnotic modification of pain. Note that there is little reason to suppose that firing in A-delta and C fibers is modified by this procedure. If pain can be modified hypnotically, this means that certain patterns of input by way of the visual and auditory modalities can modify pain.

Hypnotic reduction of pain

Several experimenters have attempted to reduce experimentally induced pain through hypnosis. Most of the studies indicate that (1) nontrance suggestions have an effect comparable to that of suggestions under hypnotic trance, and (2) verbal reports of pain are reduced by these variables, but physiological measures are not. The fact that nontrance suggestions have an effect resembling hypnotic suggestions does not mean that hypnosis is ineffective. It means that both techniques are effective. This is a good example of an effective pain-reducing technique that is regarded, because of tradition, as almost "fraudulent."

Barber and Hahn (1962) found that both waking-imagined and hypnotically suggested pain reduction lowered pain responses to cold pressor stimulation. Response measures included verbal reports, breathing changes, modifications of muscular tension in the forehead, increased heart rate, and changes in palmar skin resistance. The physiological measures of "involuntary" response (heart rate and skin resistance) were unchanged. Sutcliffe (1961) also found that verbal reports, but not skin resistance measures, indicated reduced pain under hypnosis.

Are the subjects simply lying? This is not likely, since the verbal reports measured are usually accompanied by unspecified gestural responses that lend credence to them. These should be measured. Furthermore, the research of Hilgard (1969) provides a rather convincing case for authentic pain reduction under hypnosis. It is also methodologically interesting.

Hilgard's research on hypnotic pain reduction

Hilgard's (1969) research on pain reduction exemplifies a number of the methodological principles we have discussed earlier. For one

Ernest R. Hilgard

thing, he took multiple measures of his dependent variable. Actually, verbal reports were supplemented by several physiological measures, but he presented data for only one of the physiological measures—systolic blood pressure. Secondly, he evaluated the generality of his findings by using two distinct means of producing pain—the cold pressor and ischemic pain, discussed earlier. Finally, he reduced the variability in his experiment by controlling for hypnotic susceptibility. Since he could not change the hypnotic susceptibility of a subject (someone may do this in the future—why not?), he devised a pretest that made it possible to predict which subjects would be susceptible to hypnosis. He regarded variability as a challenge to experimentation, and therefore was able to attain less noisy results than other experimenters.

EFFECTS OF HYPNOSIS ON A COLD-PRESSOR RESPONSE. Subjects were divided into two categories, high-hypnotizable and low-hypnotizable, according to their scores on the preliminary suggestibility measure. They served for three days. On the first day they were in the normal waking state, and on the second and third days they were in

one of two hypnotic conditions: hypnosis without suggested pain insensitivity, or hypnosis with suggested insensitivity (analgesia). Although Day 1 was always a day without hypnosis, the order of presentation of hypnosis, with or without analgesia, was randomized over the second two days. It would appear that there is a serious danger of order effects being confounded with both hypnotic conditions, since the normal state always came first. However, this problem was diminished because the subjects had worked in a previous experiment that gave them considerable experience with the cold pressor without hypnosis. They were presumably at asymptote with respect to pain response by the time they reached the present experiment. Put another way, the probability is not great that there was a sudden dip in pain responsiveness unrelated to hypnosis at the particular time when hypnosis happened to be introduced.

The verbal measure consisted in placing the subjects' pain on a scale from 0 to 10. As mentioned before, 0 meant no pain and 10 meant a pain so severe that the subject wished to withdraw. Scaling was done repeatedly at 5-sec intervals. The physiological measure was of blood pressure. Measures were taken every 10 sec from the fifth second on.

The results are shown in Figures 9.5 and 9.6. There appears to have been a tendency for hypnosis to reduce pain slightly, even when no suggestion of analgesia had been made. This is not surprising because any form of relaxation can reduce pain. Suggested analgesia produced a clear reduction in reported pain, more for high- than for low-hypnotizable subjects. However, as in the experiments described earlier, the physiological measure did not agree with the verbal one. In fact, blood pressure *rose* instead of declining with the hypnotic conditions. Hilgard suggested that since any stress can produce a rise in blood pressure, the observed rise might not indicate the pressure of felt pain so much as it did the presence of stress in general.

EFFECTS OF HYPNOSIS ON ISCHEMIC PAIN. Ischemic pain gave different results. Ischemic pain differs from cold pressor pain both with respect to speed of onset and with respect to whether the stress of cold is added to the stress of oxygen-debt pain. Both cold and pain can be stressors; therefore the suggestion that pain will not be felt leaves the cold stressor unimpaired.

Subjects in this experiment were selected from those in the cold-pressor experiment. The ones selected were those who were very capable of reducing pain under hypnosis. In this experiment, subjects could rid themselves of pain for from 18 to 45 min, and their blood pressure either did not rise or rose very little.

A needed control in studying the effects of hypnosis

Barber and his coworkers have pointed out the widespread tendency to underestimate the abilities of people who are not hypnotized (see, for

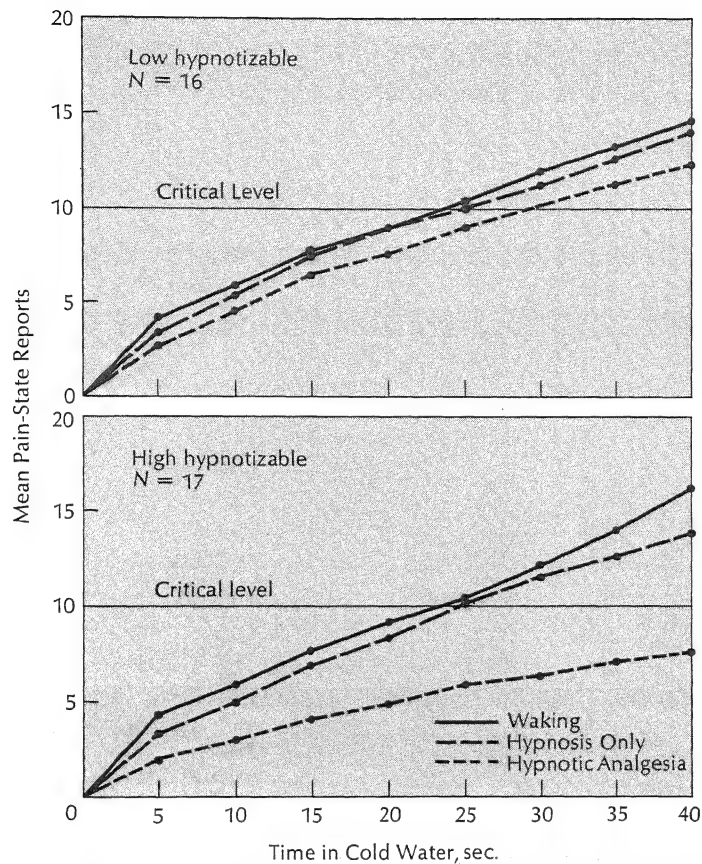



FIGURE 9.5 Effects of hypnotic suggestion on verbal reports of pain for low and high hypnotizable subjects. (From Hilgard, 1969.)

example, Barber, Spanos, & Chaves, 1974). Over a period of many years they have run control groups in which the trance-induction procedure was eliminated. Subjects were simply asked to try to perform tasks typically regarded as requiring hypnosis. They have consistently found little or no difference between the hypnotized and control subjects. Sometimes the controls are even better! For example, controls can feel as if an earlier event has been forgotten, control their dreams, change the temperature of their skin, increase visual acuity, and control allergic responses (Barber et al., 1974, Chapter 10). Obviously such control groups should be a routine component of studies on hypnosis.

COUNTERCONDITIONING AND PAIN

Stimuli that are normally quite painful can be associated with positive reinforcement, which may eliminate apparent pain responses. Pavlov

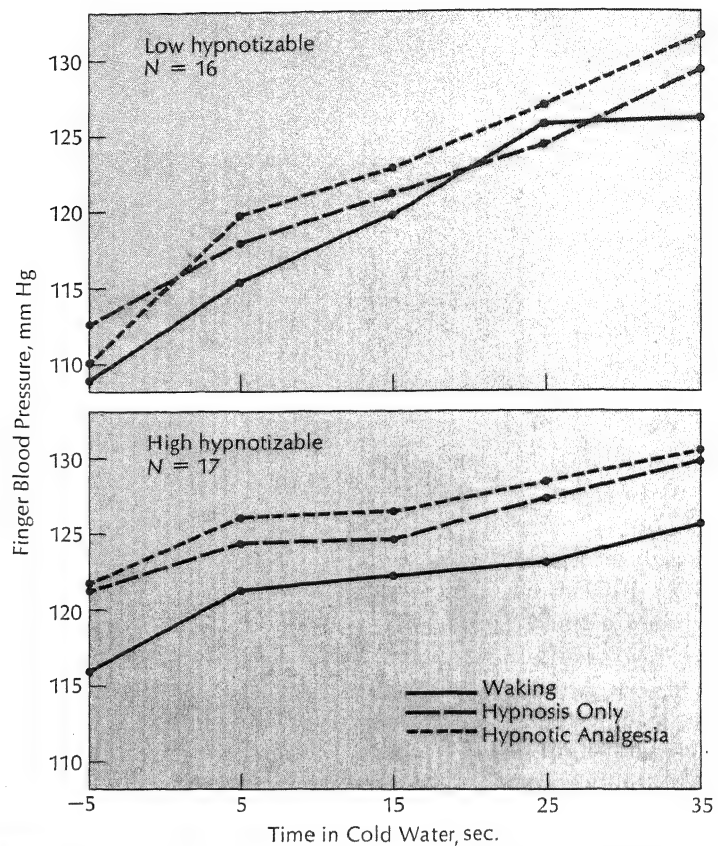



FIGURE 9.6 Effects of hypnotic suggestion on finger blood-pressure measure of pain for low and high hypnotizable subjects. (From Hilgard, 1969.)

(1928) and his students used electric shocks and pinpricks as conditioned stimuli for alimentary conditioned responses. The normal response to painful stimuli was supplanted by a nonnoxious conditioned response. His subjects failed to show even the slightest signs of pain.

More recently, Holz and Azrin (1961) showed that when pigeons received electric shocks only during sessions when food reinforcements were available, rate of responding actually increased during sessions in which punishment was programmed. Apparently shock had become an indicator of the presence of positive reinforcement for the pigeon, and the usual aversive response to such an event did not occur. This sort of work deserves further careful study with a variety of measures of pain. How many of the indicators are changed? Are the subjects still "feeling as much pain as ever" but behaving according to the training requirements? Do they continue to peck but show other

indications of pain? Pavlov's observations suggest that the other indicators *are* modified, but more systematic observations are needed.

PLACEBO EFFECTS

Patients often feel relief from various complaints when they are given what ought to be an ineffective treatment, such as a high-sounding prescription that is actually for sugar pills. It is tempting to say that they have been fooled by an ineffective drug. But if such treatments change the way the patients feel, how can they be said to be ineffective? Our tendency to regard placebo effects disparagingly is an illustration of our awesome respect for simple physiological remedies and our contempt for psychological ones. The power of placebo effects is suggested by the finding of Beecher (1960), who reported that about a third of patients being treated for postsurgical pain get as much relief from a placebo as from morphine.

Gelfand, Ullmann, and Krasner (1963), using the ultrasonic method of producing pain, gave an experimental group a placebo after a pretest of pain perception and tolerance. A control group received no pill. A significant reduction in pain was found in the experimental group as compared to the control group.

Nisbett and Schacter (1966) did an interesting study that indicates how instructions given along with a placebo can transform its effect. All subjects were given a placebo before a painful electric shock. Half of the subjects were told that the side effects of the "drug" would include arousal symptoms such as palpitation and tremor. The other half expected no symptoms. Subjects believing themselves to be in an artificial state of arousal failed to attribute their shock-created arousal to the shock, and found the shock less painful.

CONCLUSION

Many interesting experiments have been left out of this discussion of pain. For example, there is a fairly large literature on relationships between "personality" variables and pain sensitivity (see Petrie, 1967, for a summary of one block of this literature). However, the research strategies and major accomplishments in the study of pain should be clear by now. With pain, as with other types of perception, inference to structure supplemented by direct physiological measurement has been the traditional approach. This has led to a reasonably good understanding of the bodily structures underlying pain. More recently there has been a growing interest throughout the area of perception in the multiple determination of percepts. Pain is no exception. There seems to be good evidence that pain is a behavior like other behaviors, under the control of all sorts of variables, including motivational, cognitive, and affective ones.

Key Ideas Box 9.4: The physiology and psychology of pain

A great deal of work has been done on the assumption that activity in specific neural pathways was the necessary and sufficient condition for pain. Receptors were identified (probably erroneously) as the free nerve endings of the skin. Certain peripheral nerve fibers, A-delta and C fibers, seem to mediate fast and slow pain, respectively. Various spinal pathways leading to the brain have a special relation to pain. These include a "neospinothalamic" system and another system that goes through the core of the spinal cord and brain stem. There has been little success in identifying areas of the surface cortex that mediate pain, though there are some deep in the infoldings.

Recently, there has been increasing emphasis on higher psychological factors that influence pain. Pain occurs without activity in the prescribed pathways, for example, during phantom-limb pain. It fails to occur or is slight sometimes when there is activity in the pathways. There is much anecdotal evidence for this, and careful work in the laboratory supports it. For example, hypnosis alleviates pain, as does counterconditioning.

Since "psychological" variables influence pain to a greater degree than might previously have been supposed, there is a danger that the psychological approach might be regarded as somehow in competition with the physiological one. Just as some investigators will regard hypnosis or placebos as unworthy variables, others will regard their demonstrated effectiveness as indicating that psychological factors can influence and even override physiological ones. But keep in mind that hypnosis and placebo effects must also be represented by higher-level physiological mechanisms of some sort. From one point of view, it is a question of simple versus complicated physiological mechanisms. The gate-control theory of pain proposed by Melzack and Wall (1965) is in fact a physiological theory designed to take into account such things as the effectiveness of psychological variables. A clear treatment of the gate-control theory and its implications may be found in Melzack (1973).

Study Questions

1. On what grounds might it be argued that the nativism versus empiricism issue is scientifically meaningless?
2. Describe the main characteristics of "strategic questions."

3. Why is the nativism versus empiricism issue important?
4. What was Wertheimer's attitude toward the appeal to past experience as an explanation of perception? How can ecological sampling solve the problem brought up by Wertheimer?
5. Describe and evaluate the classic study by Gottschaldt on the role of past experience as a determinant of perception.
6. What is the nature and significance of von Senden's monograph on space and sight?
7. Give an example of the use of ecological variations as the independent variable in an experiment on nativism versus empiricism.
8. Explain the various techniques of quantitative and qualitative variation of experience that have been used in determining the role of experience.
9. What are the main problems involved in finding a dependent variable for research on the role of prior experience in perception?
10. Explain the major improvements in technique that have made it possible in recent years to advance our knowledge of the role of experience in perception.
11. What are the effects of dark-rearing? Support your statements with experimental evidence.
12. What are the effects of rearing in the presence of only unpatterned visual stimulation? Support your statements with experimental evidence.
13. What are the main conclusions to be drawn from visual cliff studies? What are the main limitations of this technique?
14. Explain the Fantz differential-fixation technique, and list the main findings resulting from its application.
15. Summarize the main findings from rearrangement experiments in animals.
16. What are the main features of Kohler's rearrangement experiments, and what were his main findings?
17. Describe the experiment of Held and Hein (1963) on the effects of active versus passive rearing. Point out the significance of this experiment.
18. Explain the doctrine of specific nerve energies.
19. How has the doctrine of specific nerve energies influenced the study of pain?
20. Discuss the problems entailed in defining pain.
21. List and briefly explain the main independent and dependent measures used in research on pain in the laboratory.
22. What do we know about the receptors for pain?
23. Which peripheral nerve fibers relate to pain?
24. Give evidence that pain is multiply determined.
25. Describe and evaluate Hilgard's research on the hypnotic alleviation of pain.

26. What control does T. X. Barber's research indicate as essential for studies of effects of hypnosis?
27. Give evidence that the following can influence pain:
    a. counterconditioning
    b. placebo effects

Exercise

1. Effects of Imagery on Pain. This exercise is for the more stout-hearted and healthy among you. Get some ice and place it in a pan of water large enough to hold your hand. This will provide you with a cold pressor. As a dependent variable, measure pain in response to the cold pressor. Keep it around 0 degrees Celsius (32 degrees Fahrenheit). You should check the temperature with a thermometer, but the water will tend eventually to approximate the temperature of the ice, if there is plenty of ice. Dependent measures can include verbal scaling and duration of keeping the hand in the icewater.

   As an independent variable, use the following method of alleviating pain. Before the pain-inducing stimulus is introduced, close your eyes, relax, and picture yourself on a sunny beach. Picture it in detail, the blue sky, billowy clouds, lapping waves, warm sun, and so on. Now picture your body filling with energy with every breath. Picture the energy flowing into your body and pouring into the hand you intend to leave out of the water. Concentrate your attention and all your mental energy, without straining, on that hand. Keep picturing the energy building up there as you place the test hand in the water. Stay relaxed and keep concentrating in the manner described on the hand which is outside the water.

   You will need several trials to determine whether there is a reliable effect. During control trials, do not engage in the special relaxation and concentration. Be sure to vary the order of treatment and control conditions to avoid effects of order.

10 PHYSIOLOGICAL PSYCHOLOGY Fundamentals

When Wundt founded experimental psychology at Leipzig in 1879, it was all physiological psychology. But by the middle of this century, physiological psychologists had become a small minority. In fact, to physiologize (that is, explain behavior by appeal to physiology) had become very nearly a taboo. Even now, most psychologists have little training or interest in physiological variables, and many departments of psychology content themselves with a "token" physiological man, or none at all. Yet physiological psychology is currently undergoing great revitalization.

The decline of physiological psychology seems to have been precipitated by the work of Karl Lashley. We will discuss this work in detail shortly. For now, let us say that he used experimental methods to search for the brain structures that had previously been inferred from behavior and used to explain it. He failed to find evidence that many of them existed. He started with the expectation that the workings of the brain would be easily unraveled, but he ended in virtual despair. Since Lashley was regarded as the greatest psychologist of his time, his despair was bound to lead to deep discouragement on the part of others.

What had happened was that physiological psychologists had fallen

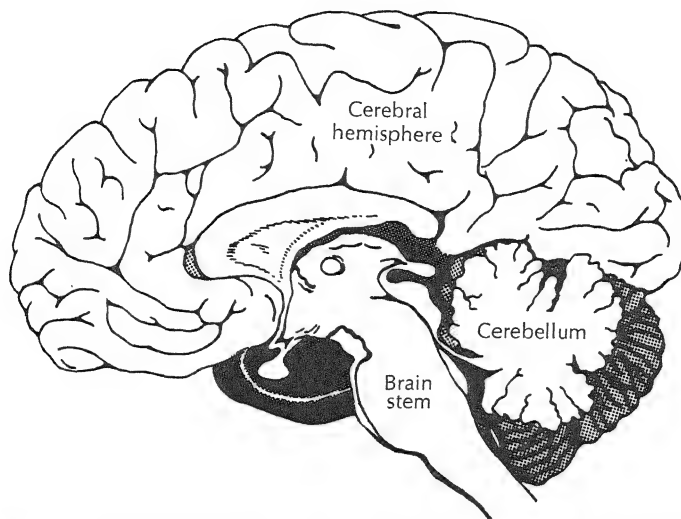



FIGURE 10.1 The brain as it would look if split lengthwise, showing the three major divisions. This drawing shows the right side of the brain. (From Wickens, & Meyer, 1961.)

function known as *classical connectionism* had been formulated, and a great deal of experimental and clinical evidence had been amassed to support it. Regions specialized for simple sensory and motor functions had been identified. Higher regions involved in organized perception and complex motor functions such as speaking had also been found. Tentative identification of the regions that organized all of this into a whole (the "highest" brain regions) had been made. Figure 10.2 illustrates the classical connectionist notion of how the brain works.

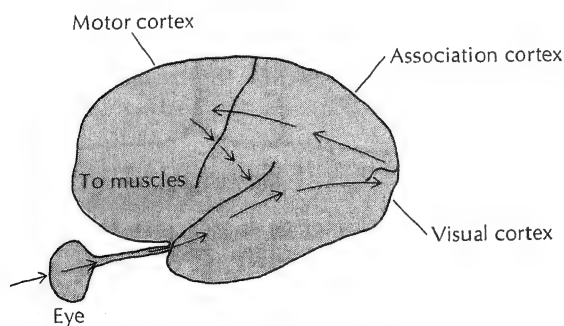



FIGURE 10.2 Highly schematic diagram of classical connectionistic view of brain function. Visual input flows from eye to visual cortex and then through association areas and on to motor region. The motor region initiates bodily movements. Learning consists of changes in the wiring diagram in the association areas.

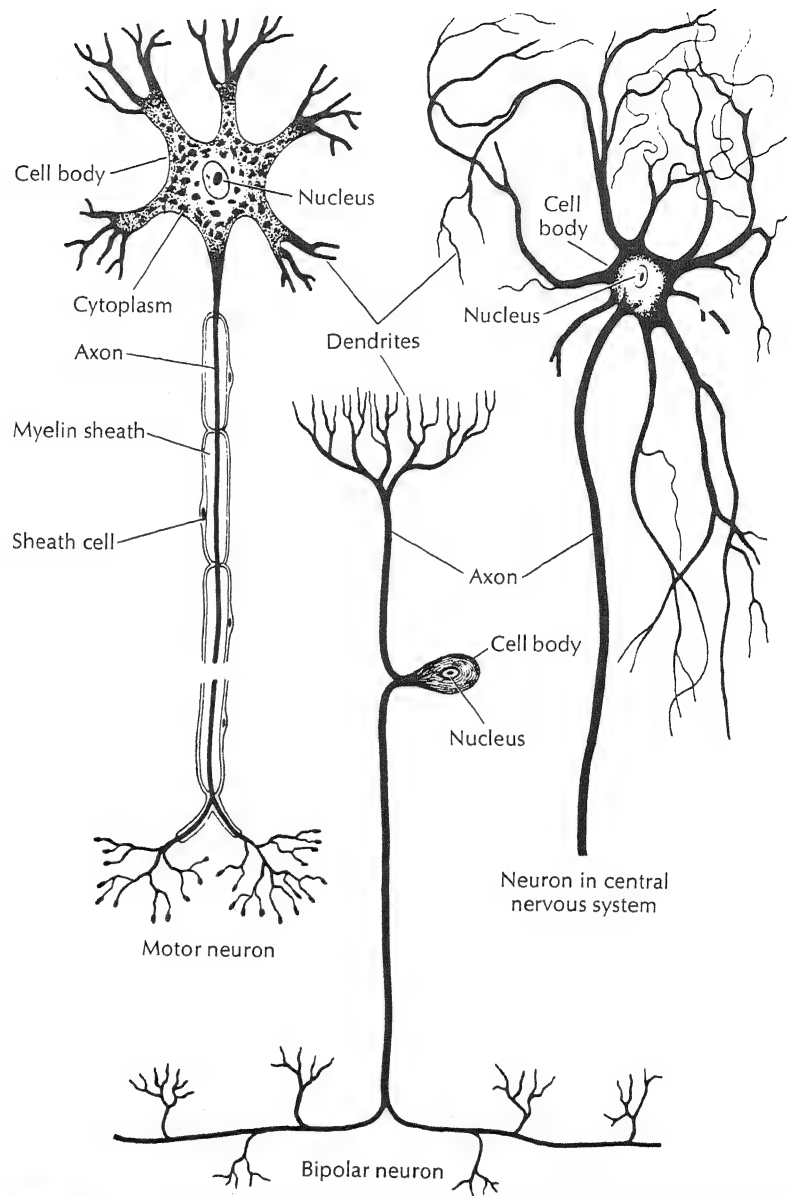



FIGURE 10.3 Three common types of neurons, all having long processes capable of transmitting information. (From Griffin, 1962.)

According to this theory, the basis of connections within and between the special functional areas was the neuron, the fundamental unit of the nervous system (see Figure 10.3). Neurons were able to communicate with each other through points, called *synapses,* where

they came very close together. Conduction occurred across synapses. An organism's response to incoming stimulation depended on the ease of conduction across synapses. Thus, learning and memory had their basis in changes in the conductive properties (or "resistance") of the synapses. Such changes in synaptic resistance presumably occurred in brain areas not occupied by simple sensory or motor functions. These were called "association areas," since they were presumed to be the basis for association between ideas, or between stimuli and responses.

At the dawning of the twentieth century, classical connectionism was very well accepted by students of cerebral functioning. The factual bases of the connectionistic point of view could hardly be called into question. But early in the twentieth century Karl Lashley set about applying experimental techniques to the problem of identifying the brain regions involved in the higher mental processes of learning and memory. His main method was to observe changes in learning and memory that occurred after he had damaged the brains of animals in various ways.

Lashley began very optimistically. He thought it would be a relatively easy task to figure out the wiring diagram of the brain and to understand how learning and memory occurred. But he ended his career many years later with no real grasp of the cerebral mechanisms underlying these higher processes. In his classic paper entitled "In search of the engram,"[1] he said that he sometimes believed that the implication of his research was that learning is impossible.

To illustrate how he became so pessimistic, let's look at his earliest work, reported in *Brain Mechanisms and Intelligence* (Lashley, 1929). Lashley did a massive series of experiments on rats in which his main independent variables were the size and the location of regions of damage to the surface layer (cortex) of the brain. (In later years he extended much of this work to primates.) His dependent variables were various measures of rate of learning of mazes of varying difficulty, as well as the degree of retention of such learning. He found that, provided the maze was a difficult one, cortical damage impaired performance. This impairment did not show up on simpler mazes. Thus there was an interaction of maze difficulty and the effects of brain damage (see Figure 10.4).

The most interesting finding was that even where cortical damage had an effect, it made no difference *where* the damage was done in the brain. Lashley proposed a *principle of equipotentiality* that says that one region is as capable as any other of performing the higher functions of learning and memory. On the other hand, the *amount* of the damage made a difference. The larger the region of damage, the greater the impairment of performance. In response to this finding he

[1]"Engram" means a memory-trace.

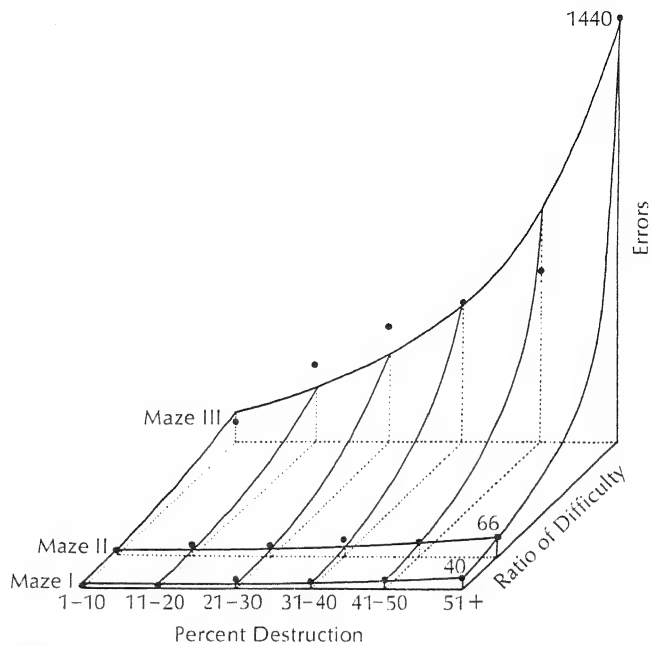



FIGURE 10.4 The relation between the extent of cerebral lesion, difficulty of the problems to be learned, and degree of retardation. The separation in depth of the curves represents the relative difficulty of the problem for normal animals; the horizontal axis of the curves—percentage destruction, and the vertical axes—number of errors made during training. (From Lashley, 1929.)

proposed a *principle of mass action.* This principle states that the size of the region damaged is an important determinant of the amount of functional loss.

Lashley's principle of equipotentiality places him squarely within the ranks of the holists. Does this mean that he denied the data on which classical connectionists based their theory? No. He acknowledged that simpler functions were localized in specific regions, such as the visual part of the brain. But he maintained that each specialized region also made a contribution to the over-all performance of learning and memory. He performed certain rather convincing experiments to show that specialized areas do something beyond their specialized function. For example, he compared the effects on maze running of cutting out the eyes of rats versus the effects of damaging the visual cortex. Damage to the visual cortex had many times the impact of removal of the eyes. Thus the visual cortex must do something more than mediating seeing.

Lashley abandoned classical connectionism and the whole notion that the brain works through operation of specific circuits that are

Key Ideas Box 10.1: Pluralism, holism, and the Lashleyan tradition

Two competing concepts of the way the brain works, *pluralism* and *holism*, have influenced the course of research. Pluralists argue that different functions are performed by different structures. Holists argue that the brain works as a whole, without special localization. During the nineteenth century, the pluralists gained ground steadily as regions specialized for movement, sensation, perception, and speech were identified. However, in the twentieth century, Karl Lashley obtained evidence that holism was valid with respect to the higher functions of learning, memory, and intelligence. He ran rats in mazes and damaged their brains in varying amounts and locations. No particular location seemed related to learning or retention of the mazes. He went on to spend a lifetime obtaining similar results with various species. He proposed several theories and principles of brain functioning. The *principle of mass action* says that it is the amount of tissue rather than its location that is important to higher processes. The *principle of equipotentiality* says that one brain area is as capable as another of performing these higher functions.

Lashley did not deny the specialization of simpler functions. He maintained that specialized brain areas, such as the visual one, also had a general role in performance of higher functions. He compared effects of removing the eyes to effects of removing visual cortex tissue, and found maze learning much worse after removal of visual cortex. This supported the idea that visual cortex functioned in some way besides direct processing of visual stimuli.

modified during learning. His research and theories still have a major influence on our views of cerebral functioning and on contemporary research.

Current attitudes toward theories of the functioning of the brain

If the brain does not work through the activation and modification of specific neural circuits, how *does* it work? Many attempts have been made to answer this question. All of them have given discouraging results. Lashley proposed that higher computations of the brain were due to the interactions of nonneuronal electrical fields in the cortex. He never worked out the mechanism of this, and later admitted that it was

hard to see how such fields could produce behavior (see Jeffress, 1951). Wolfgang Köhler tried to work out the field theory in detail (see, for example, Köhler, 1951) and put it to a number of successful tests. But other investigators found that perception and discrimination were intact despite implantations of insulators and conductors in cortical tissue. These should have disrupted the postulated fields, but they left behavior virtually undisturbed.

Other theorists, such as Hebb (1949), tried to create new concepts and models within the framework of connectionism and its notions of synaptic change. Self-organizing, random-net models that are initially wired chaotically, but change their connections when "rewarded" for the right action, have been constructed. They display many of the properties of brains. These include the ability to function purposively (for example, to find their way around obstacles in order to reach a goal) and Lashley's mass action and equipotentiality. But there is mounting evidence that the brain is not wired randomly at birth.

Most physiological psychologists work outside the framework of any broad theory. They work in the laboratory to identify the physiological variables that control behavior, seemingly with the attitude that we are just not ready for the big theory. They are theory-starved, but their findings are most interesting. Their emphasis is on applying physiological methods to psychological questions. The best way to understand today's physiological psychologists is to understand their methods and how they apply them.

Method in physiological psychology

Physiological psychology is characterized by its method, not by its content. Thus, a physiological psychologist might investigate sensory or perceptual matters; he might look into questions of intelligence, learning, or thinking; he might interest himself in emotion and motivation, or any of the other topics of interest to psychologists. He differs from other psychologists only with respect to his use of physiological method.

Since method is central to physiological psychology, I will review some of its major methods and their applications before talking about the tactics used in specific areas of interest to its practitioners.

METHODS OF IMPAIRMENT

Among the earliest methods to yield knowledge of how brains work were methods involving the impairment of the nervous system through damage. These methods were eventually extended to include the damaging of structures, such as the endocrine glands, that are not part of the nervous system. The fundamental characteristic of such methods

is that behavior under conditions of impairment is compared to behavior when the organs of behavioral control are functioning well. Experimental designs may be either of the within-subjects or of the between-subjects variety. If the impairment is irreversible, it is important to make comparisons with a separate group of subjects, since the experimenters would otherwise be using an "observe-treat-observe" pseudoexperimental design, or something of that sort.

Shallow lesioning

The word "lesion" is from a Latin word for "to wound." By one device or another the experimenter removes a part of the nervous system or deprives it of opportunities to function. Two major techniques found in the literature are *thermocautery* and *aspiration.* Thermocautery is an old technique, found only occasionally in contemporary articles, but it was used in older studies that continue to have heavy influence on the thinking of physiological psychologists. The usual thermocautery method is to heat the surface of the skull over the brain region of interest. This coagulates the brain tissue. This can be done with something as simple as a soldering iron.

Karl Lashley wanted to understand how the cortex, the surface "grey matter" of the brain, related to intelligence. He used rat subjects—many, many of them—and, by thermocautery damaged various regions to various extents, so that in the long run he had covered just about every region of cortex and a wide range of degrees of damage. He measured the ability to remember previously learned mazes and to learn new ones. His results indicated to him that there was no single region responsible for "intelligence." All of the different cortical regions made their contribution to general intelligence, even though a given region might have its own special function (such as vision or hearing) as well. Lashley's work will be treated in more detail later, but here it illustrates the important role played by the thermocautery technique in the development of physiological psychology.

The thermocautery technique has fallen into disrepute in current times. There is no exact way to regulate the spread of heat, so damage is not limited to the regions of interest. For example, if you wish to destroy cortex, the damage will inevitably include regions below the surface as well. Still, there are special instances in which thermocautery may be used. Muntz and Sutherland (1964), two highly capable experimenters, used it to study the functioning of the visual system of the rat. Rats have a visual system quite different from ours in many respects. For example, about half of the nerve fibers from each of your eyes go to the cerebral hemisphere on the same side as the eye, but the other half cross over to the opposite hemisphere. Thus, you have both an *uncrossed* and a *crossed* visual fiber system, and the two are approximately of equal size. In contrast, almost all of the fibers from a

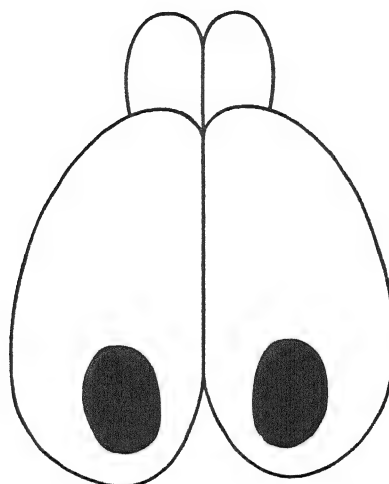



FIGURE 10.5 Striate area of a rat's brain. The blackened region designates the area that is closely related to vision in the rat. (From Maier and Schneirla, 1964.)

rat's eye cross over to the opposite hemisphere. The exact proportion depends on whether it is an albino rat or one with full pigmentation, but, in general, something like 95 percent of the fibers cross.

Are the uncrossed fibers, small as is the bundle they form, of any use to the rat? This can be answered by damaging the part of the cortex that is responsible for vision. This is called the visual cortex, and is pictured in Figure 10.5. If the visual cortex is removed on one side of the rat's brain, the animal can be restricted to either the uncrossed or the crossed visual pathway by covering one eye. If the eye opposite the damaged hemisphere is covered, the animal is restricted to the uncrossed fiber pathways, as Figure 10.5 shows. It is important that the visual cortex be completely destroyed, but damage to surrounding regions can be tolerated, so thermocautery is an acceptable technique. Muntz and Sutherland used the technique to show that fully pigmented rats can learn visual discriminations by way of the slender uncrossed visual pathway. Later studies have indicated that the same is not true of albino rats, however (Creel & Sheridan, 1966).

Most experimenters now use aspiration to make surface lesions of the cortex. A suction pump gently pulls away the tissue. If it is set at the proper level of suction, and provided the surface membranes are not permitted to intervene, the surface cortex will come off while the underlying white matter is left intact. The experimenter cannot always judge the extent of the lesion accurately, but this can be regulated by removing the treated brains and examining them microscopically after the experiment. Corrections can be made with subsequent animals.

Examples of the use of the aspiration technique are myriad. One

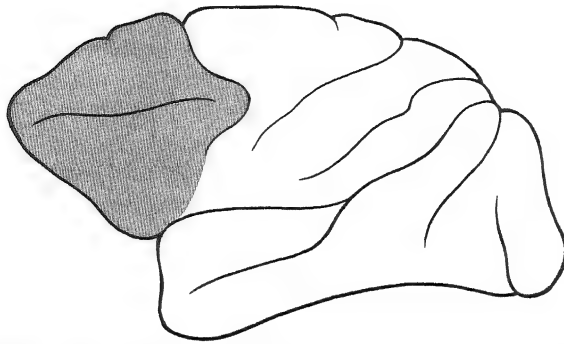



FIGURE 10.6 Side view of left hemisphere of monkey brain. The shaded area indicates frontal cortex.

chosen almost at random is the case of removal of frontal cortex in monkeys. Figure 10.6 shows a side view of the monkey brain, and the shaded area indicates the region considered to be "frontal intrinsic cortex." This is the technical name given to the part of the cortex in the front of the brain that would normally be called "higher" cortex. The brains of lower animals differ from those of higher animals particularly with respect to the expanse of "higher" cortex. This type of cortex does not have any simple sensory functions, nor is it directly involved in the control of movement. Since it is associated with advanced standing in the animal kingdom, and since it has no simpler functions, scientists have long thought that it must control some sort of "higher mental function."

Early researchers used unrefined methods, virtually scooping out the front part of the brain, but they discovered certain peculiar consequences that came to be known as the "frontal lobe syndrome." Monkeys with frontal lobe damage, like similarly damaged people, tend to show precious few gross symptoms. Such monkeys can learn difficult discrimination tasks quite readily. They lack an ability to make appropriate responses if the experimenter requires them to delay responding. For example, an experimenter prepares a tray with two foodwells covered by two identical objects. He takes a raisin (monkeys typically like raisins), waves it in front of the monkey, then places it beneath one of the objects in, say, the left foodwell. If a monkey with damage to the frontal lobe (called a "frontal monkey") gets to reach for the concealing object immediately, it will select the appropriate one. If, on the other hand, it must wait a number of seconds before getting to choose, it will behave as though it has forgotten where the raisin was placed. Normal control monkeys tolerate such delays with ease.

Deep lesioning

Thermocautery and aspiration are not generally used to make lesions below the surface of the brain. Experimenters use a *stereotaxic*

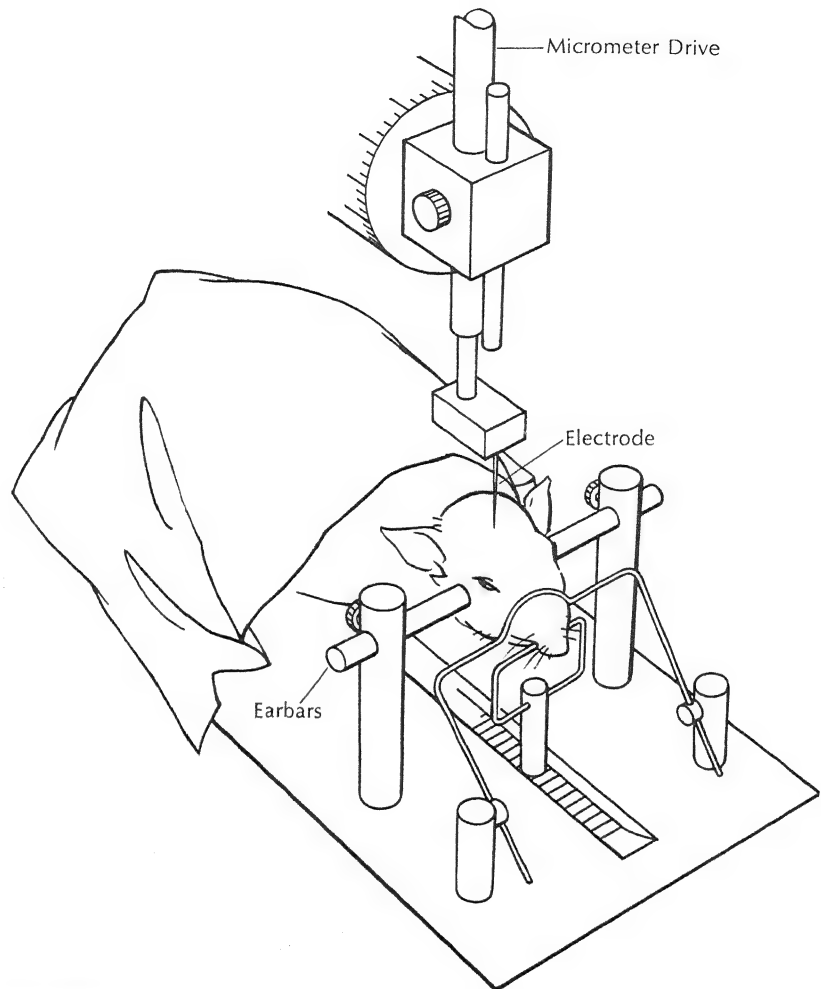



FIGURE 10.7 A stereotaxic instrument. The animal's head is placed in a standard, fixed position (in this case the ears and nose are fixed). Measured movements can be made up and down, from side to side, or forward and backward. Standard "atlases" indicate where particular brain regions are located.

instrument to do this. A stereotaxic instrument is shown in Figure 10.7. Any point in three-dimensional space can be specified by three numbers, called coordinates, one for each of the dimensions. From the front to the back of the brain is called the "anterior-posterior" dimension. From the top to the bottom of the brain is the "dorsal-ventral" dimension. From one side to the other of the brain is the "lateral" dimension. The brains of commonly used laboratory animals have been mapped so that the coordinates of any given brain structure

can be found in a *stereotaxic atlas* of the brain of that particular type of animal. The stereotaxic instrument merely provides a stable, carefully quantified system for moving a wire to the exact location specified by the atlas. The animal's head is fixed in a prescribed manner in the instrument, and micrometers drive the wire to its goal. The wire is insulated against the flow of current everywhere but at its very tip. Thus, by running current through it, damage can be done to a minute spot deep in the brain while doing negligible damage to any other structures.

Typically, experimenters run a direct current into the wire. This is called the method of electrolytic lesioning. Damage occurs at the tip as the result of generation of heat, production of gaseous bubbles due to chemical reactions, and deposit of metallic ions. Still, lesions can be restricted to very small areas. Here again, the experimenter must check the accuracy of his lesion placement by removing and examining the brain after the experiment.

Electrolytic lesioning is widely used. For example, researchers have used it to remove a part of the brain known as the *ventromedial hypothalamus.* The hypothalamus is above the roof of the mouth, fairly near the center of the base of the brain. The adjective "ventromedial" simply tells which particular part of the hypothalamus is in question. In general, the hypothalamus seems heavily involved in the regulation of basic bodily functions, motivation, and emotional states.

When experimenters lesion the ventromedial hypothalamus in a rat, the rat will stagger over to its food supply while still groggy from the anesthetic, and will begin to eat voraciously. It eats and eats and eats until it becomes grossly fat. Researchers have generally believed that this part of the hypothalamus is a "satiety center," a region that lets the rat's brain know when it is full. But other interpretations are possible, and researchers are still analyzing the nature of the changes resulting from such lesions.

"Reversible" ablations

A major disadvantage of the methods just described is their irreversibility. Experimenters have sought techniques permitting temporary deactivation of brain structures. *Spreading depression* and *electroconvulsive shock* are major candidates.

Spreading depression is a wave of deactivation that spreads slowly across the cortex and leaves it unresponsive for some time. It is a physiological state related in no particular way to the depressive mood. Experimenters usually induce the neural depression by placing a piece of filter paper soaked in a 15 percent aqueous solution of potassium chloride on the surface of the brain or its overlying covering (the solution passes through the protective membranes). Many other procedures will produce the depression, but this is the most common one.

The spread will only go over cortex. If there is a rift in the cortex, it will not pass. For example, the brain has two separate cerebral hemispheres, connected by large bundles of fibers. The wave of spreading depression will not pass over the fiber bundles. Thus, induction of spreading depression can lead to a state in which all the cortex of one cerebral hemisphere is depressed, but the cortex of the opposite hemisphere is still functioning.

This ability to depress one hemisphere at a time has been used to great advantage by many investigators, notably Bures and his associates (see, for example, Bures and Buresova, 1970). For example, they trained rats with one hemisphere depressed on a spatial discrimination task, then depressed the other hemisphere and tested them on the same task. The rats did not retain what they had learned. Even when they were given an intervening period with both hemispheres active, transfer of the learned habit from one hemisphere to the other did not occur. They also trained animals to escape shock in the absence of a warning stimulus with one hemisphere depressed, and gave them Pavlovian training by pairing a stimulus with a shock while the opposite hemisphere was depressed. Then they tested to see if the animals could combine the two sources of information, one from each hemisphere, when neither hemisphere was depressed. The synthesis of the two behaviors should provide avoidance behavior, and it did.

Electroconvulsive shock is administered by delivering an electrical current across the head. For example, a household 110-volt AC current can be delivered for a very brief period by way of clips placed on the ears of rats. This produces an electrical storm in the brain. Early investigations indicated that electroconvulsive shock resulted in a loss of memory for events that occurred just before (say for about an hour before) the shock. Long-term memories were not noticeably affected.

There is a vast literature on the effects of electroconvulsive shock. Psychologists have disputed whether memory is really impaired by such shocks. The arguments in the literature are much too lengthy and complicated to be treated here; it is enough to say that there are many studies indicating that the apparent memory losses are at least sometimes due to an inability to *use* memory stores that remain intact (see, for example, Robbins & Meyer, 1970).

Sectioning of fiber pathways

The nervous system is composed of neurons (see Figure 10.3). Neurons have an input side called a dendrite, a cell body, and an output side called an axon. Structures like the cortex are composed mainly of dendrites and cell bodies. Bundles of axons running from place to place in the nervous system, and the functions of these bundles, can be studied by cutting or *sectioning* them.

A major application of this procedure has been to section the major

fiber bundles that connect the two cerebral hemispheres. These fiber bundles are given the collective name of "forebrain commissures"; the *corpus callosum* is the largest of them (at least in mammals). The surgical operation of sectioning the forebrain commissures is called a "forebrain commissurotomy," and the result is called a *split-brain preparation.* In addition to the forebrain commissurotomy, split-brain preparations usually are submitted to a modification of the visual input pathways so that one eye projects only to the cerebral hemisphere right behind it. This is called *sectioning of the optic chiasm.* We will discuss the split-brain preparation and its effects in Chapter 11.

Final remarks on impairment techniques

Impairment techniques continue to play an important role in physiological psychology. But in this difficult field there is no entirely satisfactory technique. All lesions and sections tend to create effects remote from their site. Damage to cortex may release activity of subcortical structures otherwise held in check by the now nonfunctional region. Electrolytic lesions leave spots of degenerated and metallically embedded tissue that can produce irritation of the surround (see, for example, Reynolds, 1965). Sectioning of fiber bundles may cut off information flow directly or may simply eliminate a source of facilitation or inhibition of other brain regions. Simple conclusions cannot be drawn from such complex manipulations taken alone. Multiple, converging operations are especially important in physiological psychology. Conclusions about the functions of brain regions should be based on cumulative converging evidence from many techniques.

One fallacy particularly worth mentioning is the notion that the loss of a function after damage to a brain structure indicates that the brain structure in question was the "center" for that function. There are many things wrong with that kind of conclusion. First of all, it is not easy to say that a function is gone. To say that an organism cannot perform a certain function is to accept a null statement, and such acceptances are risky (see the earlier discussion of acceptance of null statements). Further, it may be the case that the damaged structure was an important link in the chain of events leading to performance of the function, but by no means the "center" of the function. Removing the gas tank from a car stops the car, but does it remove the "center of movement" of the car?

The other side of the coin is that survival of a function after removal of a brain region does not mean that the region did not control that function prior to surgery. For example, demonstrations that virtually all of the cerebral cortex can be removed without preventing animals from developing conditioned responses of the Pavlovian variety cannot be taken to mean that cerebral cortex has no role in the normal

Key Ideas Box 10.2: Methods of impairment in physiological psychology

Main methods used in physiological psychology include *methods of impairment* in which deficiencies in functioning after damage to the brain are observed. *Shallow lesioning,* damaging the surface of the brain, is usually done by *aspiration,* or drawing away the surface cortex by suction. *Deep lesioning,* or damaging the deeper parts of the brain, is usually done by driving a wire that is insulated everywhere but at its tip into the desired location, then running a strong electric current through the wire to damage the area. This is called *electrolytic lesioning.* The wire, or electrode, is directed to its location by way of a device that permits three-dimensional placing of the electrode. This device is called a *stereotaxic instrument* (see Figure 10.7).

Removal of surface cortex is called *ablation.* Reversible ablations have been made by *spreading depression.* A solution of potassium chloride is put on the brain, causing a spread of depression and impairment of cortical activity that passes in a few hours. *Electroconvulsive shock* has also been used in this way. Electrical current is run through the brain at an intensity sufficient to cause convulsive seizures. The electrical storm produced in the brain renders it nonfunctional for a while. Sectioning of fiber pathways is also a method of impairment. Sensory, motor, or association pathways of the brain are cut, and the resulting behavioral changes measured. Techniques of impairment, though clumsy and requiring careful interpretation, have contributed greatly to experimental physiological psychology.

development of conditioned responses. It may be that other structures *can* mediate such learning, but *normally do not.*

We have learned a great deal about the functioning of brains in the last hundred years, but the difficulties have been great. Impairment techniques have been highly important, but great support has come from other techniques like those to be discussed next.

TECHNIQUES INVOLVING STIMULATION

Like the methods of impairment, methods involving stimulation go back to the earliest period of scientific analysis of brain function; and they have played a very important role in our growing understanding of the biological bases of behavior.

Electrical stimulation

Electrical stimulation has undoubtedly been the most widely used method. Its history goes back at least to the discovery in 1870 by Fritsch and Hitzig that there is a part of the surface of the brain that, when stimulated electrically, produces movements of specific body parts. Extensions of their work have been many, and we now have detailed maps even of the human brain showing that there is a systematic representation of the body on part of the brain. The representation is upside down, and the parts of the body with refined movement (hands, vocal apparatus) have a disproportionately large amount of cortex given over to their control.

Results obtained in humans are done while the subject is awake. It is rare for any stimulation of the brain surface to hurt, and it is good for neurosurgeons to make sure where they are in the brain by getting responses from their patients before making any removals of brain tissue. An area, called Broca's area, that is involved in the expression of language is located just in front of the area controlling movement (the "motor cortex"). It is useful to get movements of the body to be sure the surgical removals will not impair language. Incidentally, electrical stimulation of Broca's area causes cessation of speech, even when the patient is coaxed to go on. This is a nice experimental method of verifying that it is a region involved in the control of speech.

The stereotaxic instrument can also permit *deep stimulation* of structures. The method is identical to that already described for deep lesioning, except that the current levels are much milder and only activate the structures instead of destroying them. (However, care must be taken to make sure the levels do not reach the destructive level.) We know that stimulation of deep structures can produce a wide array of behavioral changes. For example, electrical stimulation of the side of the hypothalamus, called the "lateral hypothalamic region," produces mouse-killing behavior in laboratory rats otherwise disinclined to do such violence. This brain region, as well as a large number of others, produces self-stimulation. This means that, if the organism is given control of the switch that regulates the current flow in its brain, it will turn it on. If it is given a brief pulse every time it throws the switch (typically the bar of an operant chamber serves as switch), it will press very rapidly and for astonishingly long periods of time.

Electrical stimulation can also be used to locate the course of fiber pathways of the brain.

A problem with electrical stimulation is that it may activate brain structures in an abnormal fashion. This may limit its usefulness in telling us about normal brain function. For example, we know that there are both excitatory and inhibitory fibers in the nervous system. Excitatory fiber endings help the cell bodies they reach to be activated, whereas inhibitory fibers result in lessening of the tendency of the

receiving cells to respond. Electrical stimulation will likely activate both excitatory and inhibitory fibers at once.

Chemical stimulation

Better than electrical stimulation in some respects is chemical stimulation of the brain. Much has been made of the well-known fact that the brain produces electrical activity, but we know that its basic mechanisms of action are chemical. The point where a neuron makes contact with another neuron is called a *synapse.* There is a *synaptic gap* between the two neurons, and it is generally bridged by a minute amount of some chemical substance, called a *"synaptic transmitter."* There are many different chemicals known or strongly suspected to be involved in transmission across synapses, including chemicals such as *acetylcholine, norepinephrine, dopamine,* and *serotonin.* Studies of chemical stimulation commonly use these or other chemicals known to be relevant to the body's functioning, such as hormones or electrolytes.

Chemical stimuli may be placed grossly into parts of the nervous system in a liquid medium. The brain is hollow, having several chambers filled with a fluid known as *cerebrospinal fluid.* These chambers we called *ventricles.* Andersonn and his associates put salt solution into the ventricle near the hypothalamus and were able to produce immediate drinking of water. Similarly, Fisher and his associates placed sex hormones in the same ventricle and noted the production of maternal nesting behavior in male rats and of male mounting behavior in females.

A more refined technique is to run a slender tube into a region of interest and then slip down a small crystal of a chemical that can dissolve in a very tiny area. This stimulation is much better localized than that obtained by putting chemicals in the ventricles. One interesting finding coming from this technique is that stimulation of the exact same spot may produce quite different behaviors, depending on the nature of the chemical used. For example, carbachol, a chemical much like the natural synaptic transmitter acetylcholine, produces drinking behavior when placed in a number of known brain regions. Stimulation with a chemical like norepinephrine in the very same spot produces eating.

Chemical stimulation of the brain with crystalline implants is one of the most refined techniques we have, but it still runs into complexities. For example, we know that, at least in some cases, the same transmitter substance produces excitatory effects when placed on one part of a receiving neuron and inhibitory effects when placed on another part. The wiring of the nervous system is so minutely detailed that even something like a chemical crystal is something of a blockbuster.

Not all chemical stimulation is done directly in the brain. In many cases chemicals are introduced into the body, either into the body

cavity or "peritoneum" (this is called *intraperitoneal injection*), into the veins *(intravenous injection)*, into the muscles *(intramuscular injection)*, or under the skin *(subcutaneous injection)*. The study done by Schacter and Singer (1962), in which epinephrine was injected into human subjects who were then tested for their susceptibility to anger- or euphoria-inducing situations, utilized intramuscular systemic injection. A study assessing the effects of intraperitoneal injection of the drug dl-amphetamine on brain-damaged rats was done by Braun, Meyer, and Meyer (1966). They showed that some of the impairment due to the brain damage was reduced by the drug.

RECORDING

We have learned about brains not just from modifying their functioning but also by recording their activity. Any number of manifestations of the activity of the brain can be measured. There are chemical changes, and a few investigators have measured them. For example, Krech and his associates have found that there are changes in the levels of acetylcholine in the brains of rats raised in enriched environments. It is even possible to record changes of brain temperature. There was already a literature on this when, at the turn of the century, William James wrote his *Principles of Psychology*. But the most popular method of recording brain activity has been electrical.

Methods of electrical recording are divided broadly into *macroelectrode* and *microelectrode* techniques, depending on the size of the lead used to pick up the signal. A gross electrode is called a macroelectrode, and is used in the typical measurement of the electroencephalogram (EEG). The research done on control of brainwaves through biofeedback techniques uses macroelectrodes to good advantage. A technique called the *evoked potential technique* typically combines recording with macroelectrodes and stimulation of the body. For example, Marshall, Talbott, and Ades (1943) flashed lights in various parts of the visual fields of monkeys while taking macroelectrode recordings from the cortex at the back of the brain, which is known to be involved in vision. They were able to get a fairly detailed map of the projection of the visual field on the brain. The visual cortex, like the motor cortex, has a very systematic relationship to the outside world.

Today, investigators working with the evoked potential technique are likely to use a *computer of average transients,* better known as a *CAT* computer, to get the average of the electrical responses to many stimulations. This computer produces an averaged curve from, say, 100 stimulation trials. There is also an instrument known as an *X–Y plotter* that will draw the graph on paper automatically and with precision. The CAT computer is deemed necessary because individual recordings may blend with experimenter bias to produce spurious results.

Key Ideas Box 10.3: Methods of stimulation and recording in physiological psychology

Electrical stimulation has long been used to identify functions of the nervous system. It may be done at the surface, where it may produce movements, sensations, or modifications of higher functions. It may also, with the help of stereotaxic implantation, be done in the deep parts of the brain where there are such things as "pleasure centers." *Chemical stimulation* may be done either systemically (into the whole body) or specifically into parts of the brain. A major advantage of chemical stimulation over electrical is that transmission in the nervous system is largely chemical, and chemicals that mimic natural activities can be used. Electrical stimulation activates the system abnormally.

Recording is usually done electrically. It is done either from various *bodily surfaces* to record such things as action of the heart (EKG), muscles (EMG), and skin (GSR), or *directly off the brain or scalp* to record brain activity (EEG). Until recently, large *"macroelectrodes"* were used almost exclusively. These record the results of large pools of units. *Microelectrodes* are now in wide use. Mostly we use *extracellular* microelectrodes that pick up potentials from nearby individual neurons. For more detailed work, *intracellular* microelectrodes pierce the neuron and tell what goes on inside it.

Whole-body chemical recording may be done by measuring chemicals in the bloodstream or their by-products in the excreta. These often reflect psychological states.

A microelectrode is a very tiny electrode, often made by carefully drawing out heated glass tubing and taking the fine wisps at the end as the shell of the electrode. Actually microelectrodes are so small that they are less thick than the wavelength of visible light, so they are too small to see with a light microscope. They are filled with conducting solution and placed in the brain with an especially stable stereotaxic instrument. The main advantage of microelectrodes is that they permit *single-unit recording.* This means that the activity of a single neuron can be observed. Single units may be either *intracellular* (within the cell) or *extracellular* (outside the cell). Intracellular recording is very difficult, since it requires penetration of the cellular wall without grave impairment of either the electrode or the cell. It is used primarily by physiologists who need special information about the electrical activity across the cellular membrane. Most work involves the extracellular method of recording. Examples of this method have been given earlier

in the section on perception, where the results of microelectrode recording have proved so enlightening.

Both electrical and chemical recording can also be done on the whole body or some organ other than the brain. Examples include recordings of the electrical resistance of the skin (Galvanic Skin Response or GSR), electrical activity of the muscles (Electromyographic recording or EMG) and electrical activity of the heart (Electrocardiogram or EKG). Chemical recording methods typically either take samples directly from the blood or assess the level of chemicals of interest by examining the concentrations of their byproducts in the excreta. An example of chemical recording would be the measurement of levels of output of certain hormones from the outer surface of the adrenal glands (17-hydroxycorticosteroids) as a function of emotional arousal. Mason (1972) has done a fine job of summarizing some important research using techniques of chemical recording.

Study Questions

1. Explain the relationship between the work of Karl Lashley and the subsequent rejection of physiological psychology.
2. Explain the difference between pluralism and holism.
3. Recount the history of research that created classical connectionism.
4. Explain the following:
   a. neurons
   b. synapses
   c. association areas
5. Describe the independent and dependent variables in Lashley's maze studies and give his major findings.
6. Explain the following:
   a. mass action
   b. equipotentiality
   c. vicarious functioning
   d. field theory
7. Explain the following and give an example of their use:
   a. shallow lesioning
   b. deep lesioning
   c. "reversible" ablations
   d. sectioning of fiber pathways
8. Give a critical evaluation of techniques involving impairment of the nervous system.
9. Explain the following and give an example of their use:
   a. surface electrical stimulation
   b. deep electrical stimulation

c. deep chemical stimulation
    d. systemic chemical stimulation
10. What are the relative advantages and disadvantages of chemical versus electrical stimulation?
11. Explain the following:
    a. microelectrode
    b. macroelectrode
    e. evoked potential technique
    d. computer of average transients
    e. X–Y plotter
12. What are the differences in method and use between extracellular and intracellular single-unit recording?
13. What are:
    a. Galvanic Skin Response
    b. Electromyograph
    c. Electroencephalogram
14. How do whole-body methods of chemical recording work?

11 PHYSIOLOGICAL PSYCHOLOGY Selected Research

The split brain

Jim L. had an outstanding record for his military performance in the Korean war. He worked his way from Private First Class to Captain. But his military career came to an abrupt halt in a freak accident. A Jeep turned over on him, producing massive injuries, including damage to the brain. He survived, but was gravely disabled. He regularly had massive convulsive seizures as a result of the accident, and eventually they became so frequent that they threatened his life.

At a time when it seemed unlikely he could recover, physicians proposed that an experimental operation be done on him. Jim was in no condition to make a judgment, so Dr. Gordon, Jim's neurologist, spoke to his wife, Helen.

"The operation involves cutting of the largest bundle of fibers in the brain—the corpus callosum. You see, Jim's seizures are started by a small region of damage left over from his accident. The damaged tissue creates abnormal electrical activity that eventually causes a massive storm of discharge from the neural tissue. We see this as the spread of the seizure over his body. The seizure starts in one half of the brain, but it spreads to the other half. We think the corpus callosum plays a role in this spread. It may also increase the likelihood of seizures in other ways. We really haven't got a full grasp of how it works. But a lot of people have improved dramatically after undergoing cutting of the corpus callosum.

The operation is still in the experimental stage. There may be long-range effects we don't know about. Keep this in mind in making your decision. But, frankly, I don't think Jim has much of a chance without this operation."

It was a grim decision for Helen to make. But she found it far less difficult than many she had been forced to make in the past. After all, there really wasn't much of an alternative. She gave her consent.

A few months later, Jim and Helen were settling in to a new and fairly comfortable life style. Jim was no longer having seizures. He still had to take it a little easy, spent most of his time reading or watching television. But he seemed better every day. They were both very glad that Helen had decided to risk the operation.

Helen's mother called long distance one day and, of course, they discussed Jim's health. "Thank God he's well, mother," she said. "Of course he has his limitations, but it's like having the old Jim back. He hasn't been this way for years."

Helen finished her conversation and went into the family room to find Jim. She saw him sitting on the sofa struggling to read the newspaper. He picked it up with one hand, and the other hand reached over and put it back down. Again he tried to pick it up. The other hand pulled it down. After several tries he managed to get the paper in front of him and began to read.

Helen watched quietly from the hallway. The doctors had told her there would be such side effects. Jim would seem normal most of the time, but once in a while there might be conflict between the two halves of his body. After all, his brain was split! She had accepted this, but there was a lump in her throat as she watched. She decided to talk to Jim later.

Though the corpus callosum is the largest bundle of fibers in the human brain, no function had been identified for it scientifically until the early 1950s. Ronald Myers, working with Roger Sperry, gave the first clear demonstration of a function for the corpus callosum. He cut this great bundle of fibers in cats, and he also split the optic chiasm (see Figure 11.1). Due to the splitting of the optic chiasm, input from a given eye goes only to the cerebral hemisphere on the same side as that eye. Thus, the visual input of the left eye is limited to the left hemisphere and the sensory input from the right eye is limited to the right hemisphere. An animal with the corpus callosum cut and with sensory input restricted to one hemisphere at a time is called a "split-brained preparation."

Myers and Sperry trained split-brained cats to discriminate patterns for food reward. One eye was covered during discrimination training. After learning had taken place with one eye covered, the blinder was switched to the other eye, and further training was given. When the cats were shifted to the second eye they did not seem to remember what

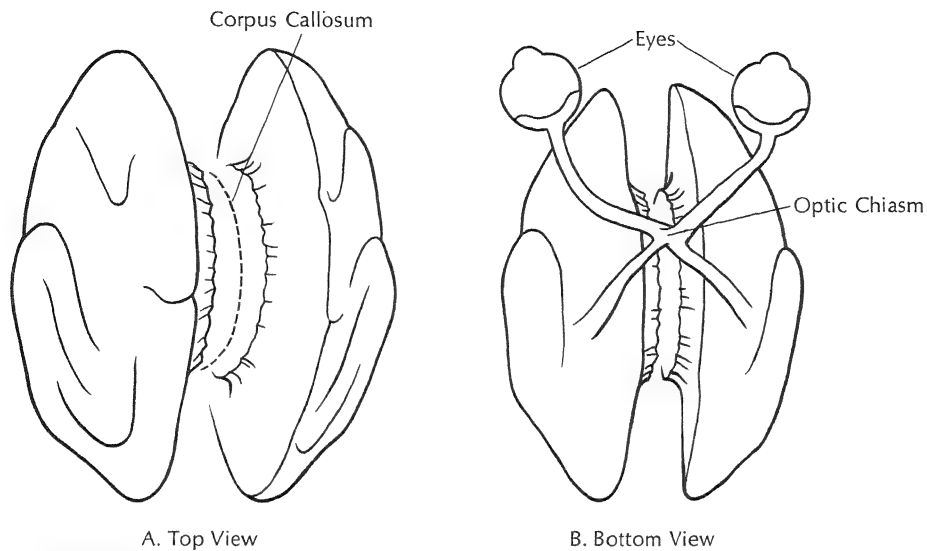



FIGURE 11.1 The split brain. Corpus callosum and optic chiasm are both split, restricting information to one half of the brain. (The diagrams are somewhat schematic.)

they had just learned. They had to relearn with the second eye in much the same way as they had learned originally, with the first eye (see Figure 11.2). The splitting of the brain seemed to create two separate memory stores, one for each hemisphere.

The finding of Myers and Sperry was very important. (See summary in Sperry, 1968, also reprinted in Sheridan, 1972.) For centuries people had speculated on the functions of the corpus callosum, but finally someone had *demonstrated* one of its functions. A second implication

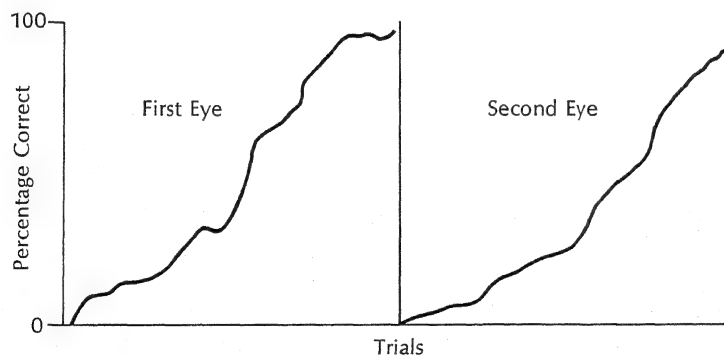



FIGURE 11.2 Idealized graph of performance of split-brained animal on interocular transfer task. The animal learns by way of the first eye, shows no retention when shifted to the second eye, and must relearn by way of the second eye.

of their finding was that memory could, to some extent, be localized. In the context of the tradition originated by Karl Lashley, this implication was highly important. A third implication was that the split-brained subject might provide an excellent baseline preparation for studying the functions of cortical regions. One hemisphere could be damaged while the other one served as a control. A final implication was the medical one. Soon after their work was published, Sperry began collaboration with neurosurgeons who helped people with severe epilepsy by splitting their brains.

The finding of Myers and Sperry is highly reliable. It has been replicated many times. In fact, cutting of the corpus callosum is what we have called a "whopper variable." But is the finding general? For example, does it apply to other species? It is—and does. Such work has been done on many species other than cats. These include rats (Sheridan, 1965; Sheridan & Shrout, 1965), rabbits (Bianki, 1959), guinea pigs (Petre & Sheridan, 1966; Lynch & Sheridan, 1970; Levinson, 1972), monkeys (Downer, 1962), chimpanzees (Black & Myers, 1964), and humans (Sperry, 1968).

The finding also applies to functions other than memory. Downer showed that emotional responsiveness could be limited to a given hemisphere by splitting the brain and then damaging the amygdala (a structure buried in the temporal lobe) on one side. Damage to the amygdala on both sides at the same time can cause a *Klüver-Bucy syndrome,* which includes such things as placidity, increased sexuality, distractibility, and a tendency to mouth objects whether they are edible or not. Downer found that certain aspects of the Klüver-Bucy syndrome could be noted in his specially prepared split-brained monkeys only when the eye on the side of the amygdaloid damage was open. For example, a monkey might be aggressive with one eye open and placid with the other open. It is as though two different personalities existed in the same animal.

Unfortunately, there are important limitations on the generality of the original findings on split-brained animals. For one thing, the results depend on the type of stimulus used. Meikle and Sechzer (1960) showed that split-brained cats can transfer light-dark discriminations from one eye to the other even when they cannot transfer pattern discriminations. More importantly, Sechzer (1964) showed that even pattern discriminations transfer from one eye to the other in split-brained cats provided they are motivated by shock. If transfer can occur under these conditions, we are not safe in using split-brained preparations as a baseline preparation for most purposes. When the second hemisphere is used as a control, information might have leaked over to it.

Since Sechzer's finding is so important, let's look at it more closely. The design involved repeated measures on the same subjects. Each

subject was trained once under shock and once under food motivation. This sort of design is commonly used when a great deal of effort has to be invested in preparing experimental animals. The split-brain operation is not easy to do, and many animals do not survive it. Thus, it is not a good idea to use the larger number of subjects required by a completely randomized design.

So, with good reason, Sechzer used a relatively small number of operated subjects (six). Two pattern-discrimination problems were given to each animal. The second discrimination problem was always the reversal of the first. Thus, a given cat might have to learn to choose horizontal stripes in preference to vertical for reward. The second problem would require that the cat choose vertical instead of horizontal stripes. We call such tasks, when the correct and incorrect cues are interchanged, *reversal tasks.* Reversal tasks are commonly more difficult than the original task.

Since the design involves repeated measures on the subjects, something must be done about possible effects of order. Normally we would counterbalance so that half the subjects received food incentive first and half were motivated by electric shock first. Then each subject would be given a *"crossover"* to the other treatment condition. Actually, Sechzer gave five of the six subjects the shock-avoidance procedure first. This may have been due to the use of randomization, but randomization only equalizes effects in the long run, not necessarily for small numbers of subjects. Thus, a very great burden of scientific significance is placed on the single cat that had the reversal task under the shock-avoidance procedure. Incidentally, this cat showed the poorest transfer of all the animals for the shock avoidance task.

Sechzer's results seem rather convincing in spite of this flaw of design. But since it is such an important study, it really should be replicated, either directly or systematically. This is especially important since another investigator (Madjkowski, 1967), using a somewhat different procedure, failed to get transfer in shock-motivated split-brained cats, and since rats show very little transfer under shock motivation (Sheridan, 1965a; Sheridan & Shrout, 1965).

BEHAVIORAL CHARACTERISTICS OF HUMANS WITH SPLIT BRAINS

Some of the most interesting work in the split-brain area comes from careful research on human subjects whose brains have been split. The pattern of behavior described for Jim is typical. There are no gross signs of abnormality, except for occasional conflict between the two halves of the body.

In order to observe the peculiarities of split-brained humans, we have to observe them under special circumstances. Something akin to

the procedure used with split-brained lower animals can be done in humans without splitting the optic chiasm. The right half of the human visual field projects to the left cerebral hemisphere, and the left half of the visual field projects to the right hemisphere (see Figure 11.3). By asking a person to look straight ahead and presenting stimuli well to the right of center, input can be limited to the left hemisphere. If stimuli are to the left of center, input is limited to the right hemisphere.

> Jim agreed to be an experimental subject. The experimenter, Dr. Hellmuth, said, "I am going to present a series of pictures to you. Your task is to identify them." She presented a picture of a star to the left hemisphere. "What do you see, Jim?" she asked.
>
> "That's easy," said Jim, "a star."
>
> "Very good. Now tell me what you see this time." Dr. Hellmuth presented a dollar sign to the right hemisphere.
>
> Jim hesitated. "Nothing . . . there's nothing there. Uh, maybe a flash of light, but no real picture."

Why did Jim fail to see what was presented to his right hemisphere? Interestingly enough, there is evidence that he really *did* see it. The right hemisphere is quite capable of seeing. But Dr. Hellmuth asked him to *tell* what he saw. In most people, the left hemisphere talks, the right hemisphere does not. But the left hemisphere *did not* see anything. So the answer was accurate for the left hemisphere. By using a nonverbal response, such as pointing, we can show that Jim's right hemisphere really can identify visual stimuli.

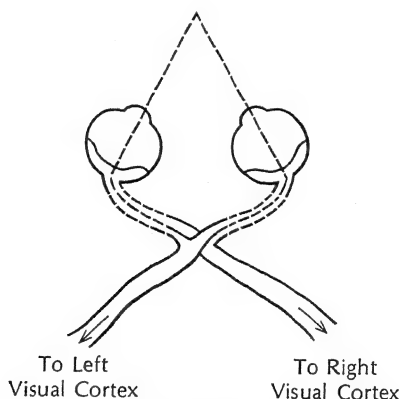



FIGURE 11.3 Things to the right of the fixation point project to the left visual cortex, and things to the left of the fixation point project to the right visual cortex. This is because the half of the retina toward the nose projects to the cerebral hemisphere opposite that eye and the other half of the retina projects to the hemisphere on the same side as that eye. By asking a human subject to look straight ahead, input can be limited to either hemisphere by placing the stimulus in the appropriate visual field. This takes the place of the splitting of the optic chiasm that is done in most lower animals.

Key Ideas Box 11.1: The split brain

When the *corpus callosum* and *optic chiasm* have been sectioned, the result is called a *split-brained* subject (see Figure 11.1). There are *no gross deficits* after splitting the corpus callosum, and this resulted in failure to identify any function for this largest bundle of fibers in the brain up until very recently. Cutting the optic chiasm permits restriction of visual input to one hemisphere, merely by covering one eye. Subjects with split brains, if they learn with one eye covered, must relearn with the opposite eye (see Figure 11.2). *Memory* seems to be *limited to one side of the brain.* These findings have proved *general* from lower animals to man.

There is evidence that, *under some circumstances, information transfers* from one eye to the other (interocular transfer) despite prior splitting of the brain. For example, simple *light-dark discriminations* so transfer. There is also evidence that even pattern discriminations transfer from one eye to the other *if subjects are motivated by shock.* Not all studies confirm the latter finding, however.

Human subjects with the corpus callosum cut can have input limited to one cerebral hemisphere (see Figure 11.3). These subjects fail to transfer complex information from one side of the brain to the other. They also show *conflict between bodily halves.* The *left hemisphere* proves to be capable of speech, though the *right* is not. However, the *right hemisphere* is capable of high-level intellectual performances. In some respects it is superior to the left.

The best way to view the split-brained subject is as a mixture of two very different people in one body. The left and right hemispheres have very different talents. The left hemisphere speaks, recognizes lists of digits very well, and tends to be good at analyzing things. The right hemisphere recognizes melodies efficiently, is good at dealing with spatial relations, and has a knack for organizing things into meaningful wholes. The two hemispheres seem to make a poor marriage, but even split-brained people seem to get along well in spite of the internal differences. Let's consider one more illustration of how the split-brained human functions.

Dr. Hellmuth presented Jim with two different things at the same time. A cross was shown to the left hemisphere and a circle to the right. "Jim, I would like you to draw what you saw. But I want you to use your *left* hand."

Jim is right-handed, but he managed to make the drawing. Dr.

> Hellmuth looked at it. Jim had drawn a circle. This is because the circle went into the right hemisphere, and the left hand is controlled by the right hemisphere.
>
> "That's fine, Jim," said Dr. Hellmuth. "Now please tell me what you just drew."
>
> Unhesitatingly, Jim answered, "A cross."

Once again, the verbal hemisphere had answered according to its own input.

Many other fascinating experiments have been done on split-brained subjects. They not only teach us about the nature of the cerebral hemispheres and their interconnections. They reveal previously unknown facts about the nature of man. A further account of them may be found in an article by Sperry (1968).

Sleep and dreaming

Sleep and dreaming are of great interest in their own right, but their study has a more general significance. It bears on the key problem of how to study inner psychological events not readily measured by ordinary means.

In graduate school I worked on a research project with the goal of devising a fully automated apparatus for measuring the learning abilities of monkeys. The apparatus was devastatingly complicated, and the plan of it existed only in the minds of a few of the senior researchers in the laboratory. Things often went haywire, and the beginners among us had to worry whether any given failure of the machine was really due to a flaw in the apparatus or to a blunder on our part. We had all, at one time or another, faced the embarrassment of calling in one of the designers at an odd hour only to find that we had failed to insert a plug or throw a switch.

Once I remember forgetting my own fear of looking foolish long enough to notice that what went on during troubleshooting exemplified the plight of behavioristic psychology. One of the designers of the apparatus had been called in because of some malfunction. He walked up to the apparatus and looked at it long and hard. He seemed not to move at all. He must have been immobile for a full five minutes. Suddenly he walked over and reconnected a loose wire. He mumbled "It'll work now," and walked out.

From the point of view of the psychology I was learning, there had been a stimulus, the complex apparatus, and a response, reconnecting the wire. How very simple! But of course, a great deal had taken place in between. We just could not see it.

The activities taking place when the troubleshooter appeared immobile had been acknowledged by behaviorists. They were *covert* re-

sponses, not observed but deduced from external behavior. How much better it would be to measure them! Today a great deal of work is done on the measurement of covert behavior. This is commonly done by the use of physiological measuring devices. It is typical for more than one measure of a given "private event" to be taken. This is, of course, the use of multiple, converging measurement operations.

One of the major sources of our increase in emphasis on measurement of covert behavior is research on sleep and dreaming. We owe researchers in this area a great deal for having worked out methods of measuring psychological events that are not grossly observable.

A major breakthrough in the study of sleep and dreaming was initiated when investigators at the University of Chicago made continuous, night-long recordings of brain-wave patterns during sleep (Aserinsky & Kleitman, 1955; Dement & Kleitman, 1957). Actually, the electroencephalographic (EEG) recordings made it possible to monitor eye movements as well as electrical potentials from the brain. The monitoring of eye movements proved to be very important.

STAGES OF SLEEP

There are a variety of sleep stages, each having its own EEG pattern (see Figure 11.4). A sleeping person passes through these stages in a recurrent pattern during the night. Five sleep stages are usually distinguished. For our purposes, the most important distinction is

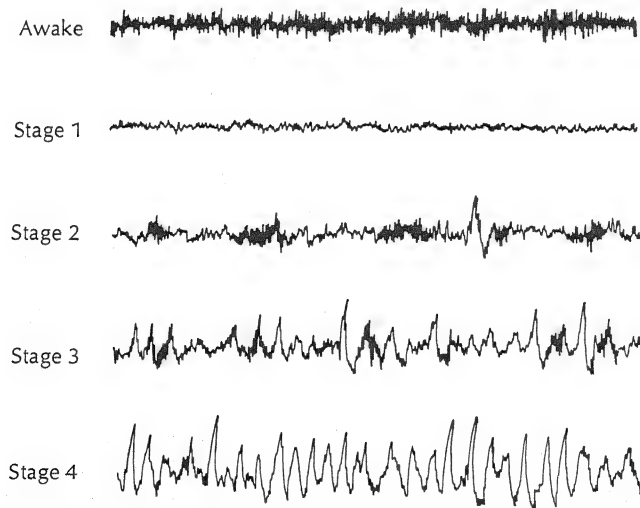



FIGURE 11.4 Typical EEG records from the waking state through the various stages of sleep. A fifth stage following Stage 4 has a pattern like that of Stage 1 and is accompanied by rapid eye movements (REM). (From Dement in Newcomb, 1965.)

between the fifth stage, during which rapid eye movements (REM sleep) occur, and the other phases taken as a whole (non-REM sleep, NREM). Detailed description can be found in Luce (1966) or Foulkes (1966). Briefly, however, the *first stage* is a phase of low-voltage, uneven, rapidly changing electrical activity. During this phase one may experience a floating sensation, drifting accompanied by idle images. One is easily awakened.

In the *second stage* the voltage is greater, and *spindle* activity occurs. Spindles are named[1] for the appearance they have as they are traced out on the EEG paper by a pen that moves up and down; the greater the voltage, the higher the pen moves on the paper. With a spindle, there is a rapid increase followed by a rapid decrease in voltage. This results in a pattern on the paper that is thin at both ends and fat in the middle. During the second stage, the eyes may appear to be slowly rolling. The subject is soundly asleep, but easy to awaken. If the eyes are held open and moderate visual stimuli are presented, the subject will not respond to them, but will often, upon awakening, testify to having been wide awake and thinking thoughts that had a vague and dreamlike quality.

During the *third stage* spindle bursts and irregular brain-wave rhythms are interspersed with large, slow waves. Voltage may be five times that found in restful waking periods. In this stage, waking the person may be relatively difficult.

The *fourth stage* has been called "oblivious sleep." A continuous train of slow waves is produced. Waking the sleeper is difficult, and upon awakening the subject may report unawareness of mental activity.

In the *fifth stage*, the sleeper begins to show an EEG pattern like that in the first stage. This is considered an ascending, rather than a descending, form of the first stage. A striking difference between ascending and descending forms of first-stage sleep is the presence of rapid eye movements (REM) in the ascending phase. Thus, the fifth stage is commonly called Stage 1 REM sleep, or simply *REM sleep*. Though in many ways the sleeper's brain activity resembles waking, it will take a great deal of noise to awaken one from this stage. On the other hand, inputs significant to one, such as one's name, may waken one easily. If awakened during this phase of sleep, the person will report vivid dreaming with visual imagery.

Regularity of the sleep cycle

On the average, we go through the various stages of sleep in repeated cycles during the night. Is there a necessary connection between the

[1]Spindles used to hold thread while weaving taper to a bulge and back again, and thus the name. Since many of us have done little weaving, a different picturesque terminology might serve us better, but we have no choice in the matter.

phases, such that Stage 1 must occur in order for Stage 2 to occur, in order for Stage 3 to occur, and so on? This would be interesting because it might indicate a strict causal connection between the phases. However, a strict connection of this type does not hold when individual data are observed. For example, a transition from Stage 4 to Stage 2 is often observed in individual subject-nights (Foulkes, 1966). Williams et al. (1964) did a study of the nightly EEG patterns of sleep in 16 medical students. Only two uniform patterns emerged. REM periods tended to occur during the last third of the night, and Stage 4, slow-wave sleep predominated during the first third of the night. Except for REM periods, subjects did not spend more than 10 min at a time in any one phase. Far from there being a strict pattern of cycles, evidence suggests that each of us has "a characteristic sleep pattern, an EEG script that is identifiable and individual, although we vary somewhat from night to night" (Luce, 1966, p. 13).

In various abnormal states the EEG patterns of sleep are markedly disturbed. Epileptics commonly have seizures during certain phases of sleep, and some patients alternate between Stages 1 and 2 without ever showing the deeper phases of sleep (Luce, 1966).

REM as an indication of dreaming

A person in REM sleep seems in many respects as though awake. But most movements are profoundly inhibited. Thus, some of the indicators suggest very deep sleep while others indicate a waking state. It is for this reason that this stage is often called "paradoxical sleep."

An obvious hypothesis is that this is a stage of dreaming. In fact, there is considerable supporting evidence for that hypothesis. Early studies showed a far higher proportion of recalled dreams when people were awakened from REM than from non-REM sleep (Aserinsky & Kleitman, 1953). Also, estimated duration of dreams proves related to REM time allowed before waking the sleeper. This and other evidence has been summarized by Dement (1965).

However, subsequent evidence indicated that some mental activity ("mentation") also takes place in non-REM sleep. Foulkes (1962) found that subjects tended to report a kind of "thinking" when awakened from non-REM sleep, whereas the reports following REM sleep were more likely to be of vivid, hallucinatory dreams. Dream-report readers instructed in the characteristics of REM and non-REM mental activity judged with 92 percent accuracy whether rapid eye movements were or were not taking place during selected dreams.

REMs correspond closely to visual imagery during dreams. However, they apparently function in some other way as well. REMs occur in the congenitally blind (Luce, 1966), in kittens reared in total darkness, and in monkeys reared without patterned visual experience (Fishbein, Schaumberg, & Weitzmann, 1965; Berger & Meier, 1965).

(See Kovlack, 1972, for further discussion of the relationship of eye movements and imagery.)

SLEEP DEPRIVATION

People have commonly spurned sleep as a "bad habit" or a "little piece of death." Think how much we could achieve if the 25 to 33 percent of the typical life spent in sleep could be used more productively. This attitude reflects the strong cultural bias we have against the nonrational modes of consciousness, and strikes me as shortsighted. It is part and parcel of a general attitude that life's sole worth comes from production. Many have studied sleep in the hope of eliminating it or, at least, greatly reducing it. But sleep is worth studying for various other reasons.

There are many situations in which sleep deprivation or disruption must occur. One of the most striking ones is in the area of aviation and space travel, but most of us have been forced to disrupt normal patterns of sleep at one time or another. Is this harmful or not? Do we injure ourselves in some permanent way as a result of sleep deprivation? What happens to performance under deprivation? Are there some conditions under which we tolerate deprivation better than under others?

The topic of sleep deprivation is a particularly interesting one from a methodological point of view. It is a topic of great human importance, one that is of interest to most of us. As Seeman and Marks (1962) pointed out some years ago, there tends to be an inverse relationship between rigor and interest value in psychological research. Interesting topics are generally studied with little methodological rigor. Topics that are studied with rigor tend to be of less apparent human impact. One reason for this is that the logistical difficulties of studying problems of immediate human interest are often severe. Sleep is no exception, yet many rigorous studies have been done in spite of the obstacles.

Studies of sleep deprivation also expose some of the harrowing problems of generality we encounter in experimental psychology. As we gain control we create an artificial situation that may lack generality. Yet we can only get knowledge of controlling variables by gaining such control. Ultimately, knowledge of controlling variables will give us generality. If we really know the causes of behavior, we can identify the situations in which such causes do or do not occur. But in the short run generality is likely to be higher if work is done in natural settings. Research on sleep deprivation has combined the use of natural and artificial settings to great advantage.

There have been a variety of reports indicating drastic consequences of prolonged suspension of sleep. As early as 1894 there was a study

showing that puppies could be killed by being kept awake. Autopsies showed they had suffered from cerebral hemorrhages. Similar findings have been obtained in more recent times (Luce, 1966).

In contrast to this, we know that some people need very little sleep. They have been called cases of "healthy insomnia" (Jones & Oswald, 1968; Meddis et al., 1973). One of these was a case of a woman who slept an average of one hour per night without naps and seemed alert and competent, with no desire for more sleep. In research on sleep deprivation, experimenters must pay close attention to differences between individuals.

Contemporary research on sleep deprivation focuses on less extreme cases than those just described. Studies usually last for a few days or, perhaps, weeks.

Total deprivation

The first thing to realize about total deprivation is that it is not easy to obtain. It is very unlikely to occur outside the laboratory because deprived subjects tend to produce brief "microsleeps." These are very short periods of sleep that are likely to appear externally as mere acts of wavering attention. It is also important to realize that the effects of deprivation depend on the influence of many other variables. One of the most important of these is the circadian cycle. Each day we go through a variety of cycles of variations in bodily temperature, hormonal output, muscular tension, alertness, and many other dimensions. Many of these cycles persist despite sleep deprivation. If subjects are tested at a "low" phase, as in the pre-dawn hours, they may show great disruption. The very next afternoon, with no intervening sleep, they may perform at almost normal levels. Longitudinal observations are much better than cross-sectional ones here.

Laboratory studies show various malfunctions with total sleep deprivation. For example, subjects fail to form memory traces (Williams & Williams, 1966). They perform poorly on simulated driving or flying tasks (Drucker et al., 1970). If they are prone to have seizures, the tendency is often aggravated. Autonomic responsiveness is impaired. For example, the galvanic skin response may not occur when the subject is stimulated (Johnson et al., 1965). Many other deficiencies have been summarized by Johnson and Naitoh (1974).

The ability of an experiment to detect effects of sleep loss depends on proper control of many different variables. For example, tasks of short duration often fail to reveal a deficit. Self-pacing in performance of a task may permit subjects to conceal deficits. Tasks of low complexity often show no deficit, as do those that provide feedback to the subjects.

Field experiments on the effects of sleep deprivation yield fewer deficits. Given the importance of such variables as those just described, this is not hard to understand. Apparently there are few cases in the

Key Ideas Box 11.2: Sleep

A major advance in research on sleep occurred when *electrophysiological measures* of it were devised. Subjects going to sleep pass from a stage of *low-voltage fast activity* of the EEG (Figure 11.4) into stages of increasing depth, characterized by *high-voltage slower waves.* However, low-voltage fast activity accompanied by *rapid eye movements (REM)* occurs after slow waves have predominated for a while. This is called REM sleep, and seems to be related to *vivid visual imagery.* However, mentation takes place in *non-REM sleep* as well. Nor are REMs perfectly correlated with visual imagery. For example, they occur in those blind from birth.

The effects of *sleep deprivation* depend on the individual and on many circumstances. Some people need very little sleep (as little as one hour per night). People can also compensate for loss of sleep through microsleeps that appear to be mere lapses of attention. For this reason, studies that demand that subjects pace themselves according to the timing of the experimenter tend to show deficits that do not appear in natural settings. Tasks of short duration and low complexity also tend to show little malfunction.

Partial sleep loss has not been studied as much as its natural occurrence might merit. A major effect of it is increasing desire to sleep. There is some evidence that tolerance for partial sleep loss may be increased by introducing it gradually. Despite early reports of a special effect of *REM sleep deprivation,* there appears to be no stage of sleep deprivation that produces more profound effects than any other. The early REM-deprivation studies have failed to be replicated.

natural environment in which we cannot conceal deficits in performance. This is probably at some physiological cost to ourselves, but it can be done.

Partial loss of sleep

Most real losses of sleep are partial. We have to cut down during crises such as final exam week. Yet there have been fewer studies of partial deprivation. One technical difficulty is that we cannot shorten the hours of sleep without also creating selective deprivation of certain stages. Many studies indicate that we spend increasing time in slow-wave sleep as hours of sleep are shortened. Thus we become deprived of REM sleep.

Studies show deficits as a result of partial sleep loss, but the absolute

loss is less important than the disruption of cycles. A major thing that happens is that people feel an increasingly strong desire to sleep. Even if they are performing adequately, they want to sleep more. There is some evidence that accomodation to shorter sleep hours can occur if the reduction is done gradually over many days (Johnson & Naitoh, 1974).

Deprivation of REM sleep

Early research (such as Dement, 1960) seemed to indicate that deprivation of REM sleep produced severe mental and emotional disturbances. A great deal was made of this. Many drugs reduce the amount of time spent in REM sleep, and some of their untoward effects were attributed to the deprivation. For example, alcohol reduces REM, and the *delirium tremens* of chronic alcoholics was speculatively attributed to that reduction. The notion that deficient REM sleep produces behavior resembling that of the psychotic has become almost a part of our folklore.

The generality of the findings was soon found to be limited. Some investigators failed to find the expected bizarre behaviors upon REM deprivation in human subjects (see Luce, 1966, for details). It is difficult to find a contemporary source to support the notion that REM deprivation produces the marked disturbances once attributed to it (Johnson & Naitoh, 1974). In fact, Johnson et al. (1974) found no differential effect on performance as a function of deprivation of different sleep stages.

Study Questions

1. What is a split-brained preparation and how does one behave?
2. What are the implications of the findings of Myers and Sperry on the behavioral abnormalities of split-brained animals?
3. How general are the findings of Myers and Sperry?
4. Under what circumstances does interocular transfer occur in split-brained cats?
5. Explain and evaluate the design of Sechzer on interocular transfer in split-brained cats as a function of type of motivation.
6. Explain the Klüver-Bucy syndrome and how it can be produced unilaterally.
7. How can human subjects be studied as split-brained preparations when their optic chiasms are not split?
8. How is the study of sleep relevant to the problems of behaviorism in dealing with covert mental activities?
9. What are the main stages of sleep?
10. What is the relationship of REM sleep to dreaming?
11. Are REMs merely concomitants of visual imagery?

12. What factors influence the outcomes of studies of sleep deprivation?
13. What are the major effects of total deprivation of sleep?
14. What are the major effects of partial sleep deprivation?
15. Does deprivation of REM sleep have any special effects?

12 HUMAN BEHAVIOR IN COMPLEX SITUATIONS Fundamentals

The need for study of complex systems

In the beginning of this book, experimental psychology was defined as a method rather than a content area. The method should, in principle, be applicable to any sort of psychological topic. Throughout the book, I have tried to show that the widespread view of experimental psychology as an accumulation of "more and more knowledge about less and less" need not be correct. There may be some experimenters who would content themselves with a deep understanding of trivia, but basic research is not trivial, and enough vitally important research has been done to demonstrate the relevance of experimental psychology to the problems of everyday living.

Yet, most of the research effort in experimental psychology has been directed toward the analysis of simple behavioral processes. It would be hard to say whether this is as it should be. There is considerable scientific precedent for it.

Science has been successful, and its methods have generally been built on the assumption that complex systems could be understood simply by understanding the elementary laws operating within them. This point of view is called *reductionism* because it reduces complex processes to elementary ones. It is called *elementarism* because it

Key Ideas Box 12.1: The need for studying complex systems

Scientific tradition places emphasis on analysis of phenomena into component parts. The mainstream of that tradition is *reductionistic.* It tends to assume that an understanding of parts of a system will automatically yield an understanding of the system as a whole. Actually only certain very simple systems work this way. Throughout science, direct study of complex systems has provided an important supplement to analysis of elementary processes. These *elementaristic* and *holistic* traditions are not mutually incompatible, but complementary. In psychology, we need a basic analysis that explains things *in principle,* but we also need an ecologically valid psychology that investigates phenomena as they naturally occur.

in their wholeness. The recent developments in earth science that have resulted from our ability to view earth systems from the holistic perspective of an orbiting body (orbited according to Newtonian principles!) is an almost poetic example of the complementary relationship of the two approaches.

If physical systems, so simple in contrast to psychological phenomena, require the study of systems in their complexity along with elementaristic analysis, there can be little doubt that complex psychological phenomena require direct study. We need both a basic psychology that explains complex phenomena "in principle" and an ecologically valid psychology that studies phenomena as they naturally occur.

The use of verbal reports in psychological research

A very common way of finding out about human behavior is to ask people to tell us about themselves. My experience with beginning experimental psychology students is that when they first start devising tactics, they go directly to the use of questionnaires or other verbal measures. One medical student recently wanted to evaluate the placebo effect of drugs as a function of the color, size, and shape of the pill. This struck me as a splendid idea until I found out that he intended to *ask* patients which of several pills they thought would be most effective for curing a given ill. From this he hoped to draw conclusions about actual effectiveness.

No such conclusions could rightly be drawn. His tactic was only valid as a means of evaluating verbal reports, and would have been appropriate if his question had focused on such reports.

The verbal report is easy to use and its use appeals to common sense. But there are so many problems entailed in the use of such reports that we must discuss them in some detail. I have broken down verbal reports into several categories, since the problems differ depending on the kind of verbal report.

MAIN CATEGORIZATIONS OF VERBAL REPORTS

Verbal reports may be *criterional* or *indicative*. If they are criterional, the verbal report is the basic datum in its own right, and is, by definition, valid. If *indicative* the verbal report is being taken as a measure of something else, another verbal or nonverbal behavior. Indicative verbal measures may not be valid. They may not be good indicators of the "other" behavior.

Verbal reports may also be categorized according to the temporal relation of the report to the event being reported. They can be *retrospective, contemporaneous,* or *prospective.*

Retrospective verbal reports

A retrospective verbal report is one in which a person is asked to report on an event that took place in the past. A widely distributed book on the female orgasm (Fisher, 1973) relied heavily on such retrospective reports. For example, women were asked to respond to the following questionnaire item:

> Please circle the answer or answers which most nearly apply to you:
>
> During the menstrual cycle I notice greatest sexual responsiveness at the following times:
>
> 1. During menstruation.
> 2. The week after menstruation ceases.
> 3. During the middle of the cycle.
> 4. The week before menstruation begins.
> 5. No differences noted during the menstrual cycle.

Fisher concluded that women are most sexually responsive during the week following menstruation.

The retrospective method is notoriously invalid. A particularly striking study of the retrospective method was that of Pyles, Stolz, and Macfarlane (1935). They compared actual data to mothers' retrospective reports on pregnancy, birth, and the early life of their children. This was done when the children were only 21 months old. The inaccuracies were striking. For example, 30 mothers whose children had suffered severe illnesses during the first year of life reported that they had no illness at all. Robbins (1963) found that parents, especially fathers, were woefully inaccurate in reporting past child-rearing prac-

tices. The errors tended to be in the direction of recommendations of "experts."

Thus we have good evidence that it is unwise to rely on retrospective verbal reports. It is surprising to see how frequently they are used in spite of this.

Contemporaneous verbal reports

Contemporaneous verbal reports are widely used by experimental psychologists. Most psychophysical experiments are done by asking people to report whether they detect a stimulus or not, whether one stimulus is greater than another, or how great is the magnitude of a given stimulus. Perceptual judgments are commonly of the contemporaneous verbal type. For example, Hilgard (1969) measured pain by asking subjects to rate the pain on a 0–10 scale, with 0 meaning no pain at all and 10 meaning a pain so severe the subject would wish to withdraw.

Contemporaneous verbal reports are usually criterional, so their validity is not in question.

Prospective verbal reports

A prospective verbal report is one requiring subjects to say what they will or would do or how they will or would react at some future time if the occasion occurred. Many verbal reports of attitude are prospective, in the sense that the verbal report is taken as a predictor of nonverbal behavior.

Certain studies have shown that prospective verbal reports sometimes fail to predict future nonverbal behavior. Milgram (1965) asked subjects how they would behave if they were asked to shock another person electrically in the setting he actually used on other subjects. There was little relationship between their estimates and the actual behavior.

Prospective verbal reports can be taken as criterional or indicative. For example, verbal reports may be as legitimate a measure of attitude as are nonverbal behaviors. If they are taken as indicative, then it is hard to say whether they are valid without directly checking their validity.

On the other hand, prospective verbal reports have sometimes been shown to be valid predictors of nonverbal behavior. For example, Rokeach (1960) asked subjects to range an array of religions in order of similarity to their own religion. Out of this, he obtained a scale of distances between religions (see Figure 12.1). He went on to examine church records from two churches in each of six denominations and found that the degree of similarity on the scale predicted migrations from one church to another.

The inadequacies of prospective verbal reports have probably been

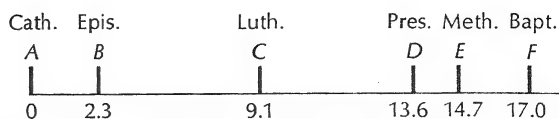



FIGURE 12.1 Relative psychological distances between religions according to Dawes (1973; based on data from Rokeach, 1960).

exaggerated in the psychological literature. A widely cited study by La Piere (1934) is often used as proof that such verbal reports do not correspond to actual behavior. La Piere traveled the western United States with a Chinese couple. Out of 251 places where they stopped to eat or sleep, they were refused service only once. Six months later, La Piere wrote to the 250 accommodating places asking whether they would serve Chinese. 50 percent of them responded; 90 percent of the respondents said "No," they would not accommodate Chinese.

On the surface, this seems to say that there is little correspondence between verbal and nonverbal behavior. But Dillehay (1973) has pointed out that the people answering the letters were probably not the same as the people who had been accommodating. The waitress in a diner or the desk clerk in a hotel are not likely to be the ones to receive and answer such a letter. Similar criticisms apply to other widely cited studies that indicate a discrepancy between prospective verbal reports and actual behaviors.

Nevertheless, there are enough valid instances of discrepancy between verbal and nonverbal behavior to make us worry a great deal

Key Ideas Box 12.2: The use of verbal reports

The main issue here is whether to use verbal or direct behavioral measures. Verbal reports are often invalid predictors of nonverbal behaviors, but they are sometimes valid.

Verbal reports may be either *criterional* or *indicative*. If criterional, they are of interest in their own right, not as predictors of anything else. In this case, they are always valid. If indicative, they are used to indicate other behaviors, and may be invalid. Verbal reports may be used to measure past (*retrospective* reports), present (*contemporaneous* reports) or future (*prospective* reports) behaviors. Retrospective verbal reports are commonly poor indicators of nonverbal behavior. Contemporaneous verbal reports are commonly valid because they are criterional. Prospective verbal reports are sometimes valid and sometimes not. Since they are of unpredictable validity, it is best to avoid using them as indicative.

about relying on the verbal reports. Until we know pretty exactly when the prospective reports do and when they do not work, it is best to avoid relying on them.

Dealing with human subjects

Psychologists have generally assumed that an experimental outcome would be uninterpretable if human subjects were aware of the experimental procedure (though there is evidence that this may not be true; see Bermant & Starr, 1972.) Consequently, they have gone to great lengths to deceive subjects. In fact, we now have a sorry state of affairs in which nearly all volunteers assume that they are being deceived when they participate in a psychological experiment.

Some investigators have proposed that the deception be eliminated and that fully informed subjects merely be asked to behave as they would if they did not know the procedure. This is called the *role-playing technique.* On the surface of it, the notion seems absurd, but it commonly works rather well. People will often produce very much the same behavior whether they are informed of the procedure or not (see, for example, O'Leary, Willis, & Tomich, 1970, or Bermant & Starr, 1973). Unfortunately, it is impossible to know when role playing will and when it will not produce results identical to those found when subjects are deceived. There is evidence that procedures involving role playing do not always give results like those involving deception (Willis & Willis, 1970), and to decide that they *do* give equivalent results involves us in a null acceptance. Since we cannot say when role playing does and when it does not give authentic results, the technique is of little value despite the remarkable ability of people to duplicate authentic effects *sometimes* or even *often* when they are playing a role. As Miller (1972) has pointed out,[1] the experimenter would always be faced ultimately with doing the real (deceptive) experiment in order to determine the generality of findings based on role playing.

Having rejected role playing as a fundamental method in psychological research, what about the use of deception? Many experimenters dislike deception because it involves a dishonest relationship between two human beings, experimenter and subject. To a degree, this is true. But it is important to look closely at what we mean by deception. Menges (1973), in an examination of the use of deception in psychological research, pointed out that at least three categories of information

[1]Role playing should be carefully distinguished from merely asking subjects what they would do in a particular situation. Miller (1972) seems to treat role playing and verbal reports as identical, and attacks the role-playing method by pointing out instances in which verbal statements of what a person would do in a given situation proved inaccurate. Getting into the role might make a great deal of difference.

may be given to the subject: inaccurate information, incomplete information, and complete information. An even more elaborate categorization could be made by considering *when* a given type of information or misinformation is given; whether before, during, or after the experiment. Furthermore, the notion of "completeness" of information is not entirely clear. Does this mean every single detail of the design? Subjects might default out of boredom or embarrassment over their inability to comprehend a complex design.

Let me describe a procedure that strikes me as fundamentally honest and workable for most experiments on human subjects. The subjects are told with emphasis that they are free to participate or not participate. Subjects are often obtained from classes in introductory psychology, where they have a requirement that they gain some sort of concrete experience in psychology. They are emphatically reassured that they can do something besides participate in an experiment, such as visit a clinic, interview an experimenter, or the like. However, they are told how important the experiment seems to be and informed of the experimenter's obligation to reciprocate, if they participate, by doing everything feasible to see to it that it is a learning experience for them. The experimenter also tells them the extent to which the details of the experiment can be revealed up to the time when it is over. They then choose whether the gains of learning something and contributing to the expansion of our knowledge of human beings are worth the losses entailed in participating without knowing whether the information during the experiment will be incomplete or even wrong.

After the experiment, the experimenters must try to live up to the other side of the bargain, immediately if possible. If it is important to prevent "leaks" to future subjects, the experimenter can take information on how to reach the participants to inform them later of important details that must not get out. Every effort should be made to avoid "using" people or pushing them around, and there should be no ill will toward people who do not participate, since we neither know nor have the right to know their private reasons for deciding as they do. My own policy is to be willing to submit myself to more than I would ever ask another person to tolerate.

This procedure strikes me as fundamentally honest and respectful, though it leaves room for withholding information at various stages in the process of doing research. It is also workable for most psychological experiments, though it creates the risk of getting unrepresentative samples. This risk is perplexing, but there may be no superior options. If a very large proportion of the available subjects participate, the risk may be quite small. If a small proportion of available subjects volunteer, a close testing of the generality of findings obtained is required. (See Logistics Box 12.1 for fuller information on the ethical principles binding psychologists who work with human subjects.)

If we agree that experimenter and subject can relate honestly and with mutual respect within the framework of at least one kind of "deception," there remain problems of more serious deception as well as more purely methodological difficulties entailed in deceptive procedures. Decisions about whether to engage in more serious deception (as when experiments are done in a natural setting without the consent of the participants) are to be taken very seriously. Peers not involved in the experiment should participate in such decisions. These decisions have to be made separately for each experiment, so I will not try to deal with them in a general way here.

The *methodological* problems involved in experiments using deception are probably more serious than most experimenters realize. We have rejected the alternative of using role playing, but must acknowledge the further difficulties entailed in using deception. Students of human behavior currently find themselves "between a rock and a hard place." In deceptive experiments it may be very difficult to be sure that the deception worked, and the study may be more a matter of game-playing between experimenter and subject than of authentic, generalizable research. We are presently in a stage of discovering how serious these problems are, and a good deal of research needs to be done before we can confidently adopt a given experimental procedure. Here, on the basis of present knowledge, I will discuss the problems of authenticity of variables and of the "demand characteristics" of experiments.

AUTHENTICITY OF MANIPULATED VARIABLES

Many of the variables of interest to those who study complex human behavior are difficult to manipulate at will. If we want to study willingness to intervene in an emergency, how are we to produce an emergency? If we wish to study willingness to obey malevolent commands, how are we to produce authentic malevolent commands?

This is a problem that should be given very careful attention by experimenters. It is not enough to have ill-trained actors produce a caricature of the real situation. In fact, even trained actors might not be able to produce the desired impact, since the simulations that are effective on stage may not work in a closer setting. Staub (1974), who wished to study the willingness of passersby to help someone under distress, went to considerable lengths to assure credibility. One technique used in the study of helping behavior involves making a tape recording of an emergency, then playing the recorded emergency while the subject is in an adjacent room. Staub pointed out the importance of having very high-quality recordings that were tested for their capacity to appear real.

Staub (1974) also went out onto the streets with some of his tests.

Logistics Box 12.1: Ethical principles of the American Psychological Association in relation to research with human subjects

The decision to undertake research should rest upon a considered judgment by the individual psychologist about how best to contribute to psychological science and to human welfare. The responsible psychologist weighs alternative directions in which personal energies and resources might be invested. Having made the decision to conduct research, psychologists must carry out their investigations with respect for the people who participate and with concern for their dignity and welfare. The Principles that follow make explicit the investigator's ethical responsibilities toward participants over the course of research, from the initial decision to pursue a study to the steps necessary to protect the confidentiality of research data. These Principles should be interpreted in terms of the context provided in the complete document offered as a supplement to these Principles.

1. In planning a study the investigator has the personal responsibility to make a careful evaluation of its ethical acceptability, taking into account these Principles for research with human beings. To the extent that this appraisal, weighing scientific and humane values, suggests a deviation from any Principle, the investigator incurs an increasingly serious obligation to seek ethical advice and to observe more stringent safeguards to protect the rights of the human research participant.
2. Responsibility for the establishment and maintenance of acceptable ethical practice in research always remains with the individual investigator. The investigator is also responsible for the ethical treatment of research participants by collaborators, assistants, students, and employees, all of whom, however, incur parallel obligations.
3. Ethical practice requires the investigator to inform the participant of all features of the research that reasonably might be expected to influence willingness to particpate and to explain all other aspects of the research about which the participant inquires. Failure to make full disclosure gives added emphasis to the investigator's responsibility to protect the welfare and dignity of the research participant.
4. Openness and honesty are essential characteristics of the relationship between investigator and research participant. When the methodological requirements of a study necessitate concealment or deception, the investigator is required to ensure the participant's understanding of the reasons for this action and to restore the quality of the relationship with the investigator.
5. Ethical research practice requires the investigator to respect the

individual's freedom to decline to participate in research or to discontinue participation at any time. The obligation to protect this freedom requires special vigilance when the investigator is in a position of power over the participant. The decision to limit this freedom increases the investigator's responsibility to protect the participant's dignity and welfare.

6. Ethically acceptable research begins with the establishment of a clear and fair agreement between the investigator and the research participant that clarifies the responsibilities of each. The investigator has the obligation to honor all promises and commitments included in that agreement.
7. The ethical investigator protects participants from physical and mental discomfort, harm, and danger. If the risk of such consequences exists, the investigator is required to inform the participant of that fact, secure consent before proceeding, and take all possible measures to minimize distress. A research procedure may not be used if it is likely to cause serious and lasting harm to participants.
8. After the data are collected, ethical practice requires the investigator to provide the participant with a full clarification of the nature of the study and to remove any misconceptions that may have arisen. Where scientific or human values justify delaying or withholding information, the investigator acquires a special responsibility to assure that there are no damaging consequences for the participant.
9. Where research procedures may result in undesirable consequences for the participant, the investigator has the responsibility to detect and remove or correct these consequences, including, where relevant, long-term aftereffects.
10. Information obtained about the research participants during the course of an investigation is confidential. When the possibility exists that others may obtain access to such information, ethical research practice requires that this possibility, together with the plans for protecting confidentiality, be explained to the participants as a part of the procedure for obtaining informed consent.

This is likely to make subjects far less prone to suspect they are part of an experiment. Unfortunately, even Staub fell into the use of an implausible situation when he had a college student go into a lower-middle-class neighborhood and feign a heart attack by grabbing his chest and collapsing. It is unlikely that a healthy young college student would be having a heart attack in the streets. Far more likely that such a healthy young man might be a practical joker or be trying to lure someone near in order to take a wallet or purse!

In a later experiment Staub (1974) corrected this error to a degree by having an overweight student play the role. In this more plausible situation, the behavior of passersby was somewhat different. More people intervened.

It is not always easy to create plausible situations, but there is little point in doing such experiments without having given this matter very careful attention.

DEMAND CHARACTERISTICS OF EXPERIMENTS

Experimenters who deal with human subjects are currently plagued by concern over the influence of demand characteristics on the outcomes of experiments. The idea of demand characteristics goes back to at least 1933 (Rosensweig, 1933, cited in Silverman & Shulman, 1970), but the major impetus for concern came from a very important paper by Orne (1962).

Orne found that human subjects were remarkably compliant with various requirements imposed on them by experimenters. He was interested in testing the influence of hypnosis on performance of various tasks, and he sought to find a task at which subjects in the waking state would balk. Such a task would act as a control for evaluating the influence of the special hypnotic relationship. But Orne found himself hard pressed to find any task subjects would refuse to do.

He had subjects add pairs of numbers taken from sheets with random numbers printed on them. There were 224 additions to be done for each sheet, and subjects were placed in a room with 2000 sheets! The experimenter assigned them the task, explaining that he would come back later. Says Orne, "Five and one-half hours later, the *experimenter* gave up!"

The task was made even more obviously meaningless by having the subjects tear up each sheet after completing the additions, but subjects still persisted. Interviews revealed that the subjects gave meaning to the task by conjecturing that the experimenter must have in mind various hypotheses that would make sense of it. For example, some subjects thought that it must be an endurance test.

Out of these and other observations, Orne derived the concept of demand characteristics. The demand characteristics of an experiment are various cues that give the subject a notion of the experimental hypothesis. Orne stressed the idea that human subjects often are controlled by the wish to comply with the experimenter's expectations. Subsequent writers have suggested that even when subjects choose not to comply with the experimenter's expectations, even when they choose to defy them, their awareness of the demands may have an important influence (Rosenberg, 1965; Page, 1971, 1973). Subjects may

suppose that the experimenter expects them to do a certain thing, but that they will look bad if they do it. So they do something else. They are said to have "evaluation apprehension" and to engage in "impression management." In order to give a good impression of themselves they may, given their notion of the experimental demands, comply or do something quite contrary to their notion of compliance. They are less likely to comply when the demand characteristics are obvious.

The influence of demand characteristics is a grave problem. In fact, we are becoming increasingly aware only very recently of just how pervasive their influence might be. Orne (1962) showed that much of the effect of perceptual deprivation occurs when a number of the demand cues are provided without the deprivation. An important role for demand characteristics has been proposed, and in some instances demonstrated, for such diverse areas as the classical conditioning of attitudes (Page, 1969, 1971), the operant conditioning of verbal behavior (Page & Lumia, 1968; Page, 1970, 1971, 1972; Patty & Page, 1973), attitude change (Silverman & Regula, 1968; Page, 1973), obedience to authority (Orne & Holland, 1968), hypnosis (Barber, 1969), and many others.

Measuring and controlling for demand characteristics

Now that we know about the influence of demand characteristics, it has become important that they be taken into account in designing, conducting, and interpreting studies involving human subjects. Much research of this kind done in the past will have to be reexamined with demand characteristics in mind. It would be dangerous to ignore demand characteristics, but let us not be tempted to use them as an explanation for observed results unless proper evidence for such effects has been obtained. It would be easy to fall into using demand characteristics as a pseudoexplanation. Remember, a functional explanation is not adequate unless at least two separate measures have been taken, one for the dependent variable and one for the independent variable. Thus, experimenters must measure demand characteristics if such cues are to be used to explain observations.

Unfortunately, adequate measures of demand characteristics are just beginning to be developed. For the most part, experimenters have been forced to rely on verbal reports from subjects. Most common has been *postexperimental questioning.* Subjects are merely asked how they perceived the situation. This may not be as simple and straightforward as it sounds. Verbal reports are always suspect. In some cases the questioning itself may provoke subjects to say that they perceived the experiment in a certain way. In particular, subjects might tell an inaccurate story if, after the fact, they perceived their behavior as putting them in a bad light. We cannot assume that whatever subjects

say is true, since they may be rationalizing or trying to create a good impression.

Page (1971) has provided evidence that the way the postexperimental interview is conducted can have a heavy influence on the results. He showed that single, open-ended questions missed many individuals who were aware of the demands of his experiments. A multiple-question technique was necessary to detect many of them. So it may be necessary to do a great deal of probing in order to identify subjects who are aware of the demand characteristics of the experiment. Unfortunately, in so doing, the risks of influencing subjects by suggestions go up.

Questioning should include inquiry about *when* the subject became aware of the demands. This may detect some of the subjects who became aware late, or even during the postexperimental inquiry. Subjects often may think it makes them look better if they say they knew what was really going on in an experiment. In some instances their behavior may seem reprehensible once they understand the real nature of the experiment, and they use the classic ego-defense mechanisms of rationalization to account for their behavior. Thus, it is hard to tell whether they really knew the nature of the experiment or not on the basis of verbal reports obtained after the fact.

A further problem with postexperimental questioning lies in what the experimenter does with the subjects who indicate that they understood the experimental hypothesis. Typically one eliminates them from the analysis of data. However, in so doing one limits the generality of one's findings, perhaps seriously, by creating a nonrandom sample of subjects. It is hard to say how subjects who do catch on to the hypothesis differ from those who do not, but it might be the case that in throwing out the aware subjects, the experimenter is eliminating the more intelligent, perceptive, sensitive subjects. If subjects are only eliminated from the group receiving the experimental treatment, this creates a confounded variable.

Pre-experimental inquiry may have the advantage over postexperimental questioning. With pre-experimental questioning, a control group of subjects is submitted to all of the experimental procedure up to the point of actually taking them through the procedure. At that point, they are asked to describe their perceptions of what is expected of them. A limitation on this method is that some discovery of demand characteristics might take place during the actual conduct of the experiment.

An interesting alternative to the pre-experimental interview is to direct subjects to behave in a manner opposite to that which they perceive as expected of them. Page (1974) showed that subjects so instructed in an experiment on the respondent conditioning of attitudes produced results opposite to those normally obtained in such

Key Ideas Box 12.3: Deception, role playing, and demand characteristics

Human subjects must be dealt with ethically (See Logistics Box 12.1 for ethical principles of the American Psychological Association). Yet most researchers assume that they can only gather valid data from human subjects by misinforming them. Subjects can, within an ethical framework, agree to participate in experiments without receiving accurate information, so the ethical problem, with proper care, can be solved. Subjects may also deceive experimenters or, at least, behave in abnormal ways because they are in an experiment. This creates a serious problem of generality. Subjects will go to great lengths to comply with what they believe to be the intentions of the experimenter (*demand characteristics* of the experiment). They will also sometimes fail to comply because they want to look good. This is *impression management.*

Some researchers have suggested full disclosure of the experimental design, simply asking subjects to *role play* as they would if they were authentically in the experiment. Results with this method are remarkably similar to those gotten with deception. However, role playing does not always reproduce the effects of the usual deceptive mode. Thus, we cannot know when it produces generalizable results.

If "deception" is to be used (and there appears to be no fully adequate alternative), great care must be taken to make the situation appear authentic. And *demand characteristics must be measured.* There are several ways to do this. *Postexperimental interviews* may be inadequate. *Pre-experimental interviews* (just before subjects actually participate) are probably better. A control group asked to *behave in a manner opposite* to what they think is expected of them is an excellent control.

experiments, indicating that the process controlling their behavior was not respondent conditioning, but compliance with demand characteristics.

Another method of dealing with demand characteristics is to ask subjects to simulate or role play in a situation lacking the independent variable (Orne, 1962). For example, subjects can be asked to act as though they were hypnotized in order to see whether there is anything unusual about subjects who are actually hypnotized. It proves to be very difficult to find any differences (Barber, 1969). The simulation

procedure may not always work, however. In certain types of experiments, the simulation procedure may be effectively identical to the real treatment procedure because subjects may suspect a trick and perceive other demand characteristics. I will discuss such a case in the treatment of obedience to authority (Wilson, O'Leary, & Tomich, 1970). In the case of hypnosis, it is possible to argue that the introduction of demands by an experimenter is itself a hypnotic induction procedure.

A second objection to the simulation procedure is that we cannot conclude from duplication of a phenomenon under a given set of conditions that those same conditions normally control behavior. There is typically "more than one way to skin a cat," and it might happen that a phenomenon can be produced either by simulation or authentically, the underlying processes being quite different.

Silverman and Shulman (1970) have argued that we might deal with demand characteristics by using indirect, unobtrusive measures in laboratory settings, thus minimizing the likelihood that demand characteristics will be perceived. A second alternative they suggest is to experiment in naturalistic settings, once again minimizing or eliminating demand characteristics. In general, increasing the authenticity of an experimental setting should reduce the likelihood that demand characteristics determine the behavior.

Probably the most important test of the role of demand characteristics in experiments is the generality of experimental findings. If experiments lead to concepts or models of human behavior that work in accounting for what goes on in the real world, we need not concern ourselves over demand characteristics. The reason for worrying about demand characteristics is that they might create results that are not ecologically valid.

Study Questions

1. Why should we study complex human behavior directly?
2. What are holism and elementarism, and how are they related?
3. What is ecological psychology?
4. Explain and evaluate the main types of verbal report.
5. Evaluate role playing as a method in psychological research.
6. Describe an ethical procedure for dealing with human subjects.
7. What are the problems entailed in using "deceptive" experimental procedures?
8. Explain the following:
   a. demand characteristics
   b. impression management
9. Explain and evaluate the main methods of measuring and controlling for demand characteristics.

Exercise

1. The Role of Experimental Demand Characteristics in Controlling Human Behavior. This experiment requires that you get others to participate. Friends and acquaintances will do. You will also need to put together copies of sheets of random numbers. The independent variable will be the act of telling participants, "This is part of an experiment." Controls will simply be asked to perform the same task without being told it is "part of an experiment." The task is to add up each successive pair of random digits and write their sum beneath them.

   What effect does being told they are "in an experiment" have on willingness to comply with this task? Afterwards, do an intensive interview on participants to see what they thought your purpose was, when they thought it, and so on. Be sure to question controls as well as experimental subjects.

13 HUMAN BEHAVIOR IN COMPLEX SITUATIONS Selected Research

Behavior and obesity

"Just looking at myself in a store window makes me feel terrible. It's gotten so I am very careful not to look by accident. It's a feeling that people have the right to hate me and hate anyone who looks as fat as me. As soon as I see myself I feel an uncontrollable burst of hatred. I just look at myself and say "I hate you, you're loathsome!"

These are the words of an overweight man. He is not atypical. Comments of other overweight people also show the broad destructive effects of obesity on their view of themselves and of others. A woman says, "Nobody wants to go out with a tub, which was my nickname. By going to a dance all I did was stand against the wall listening to music. Nobody ever asked me to dance. I made believe I didn't care. But I just thought I was a big nothing."

Another overweight man said, "Physical appearances seemed to exclude me from social activities and I resented it. I felt I was not physically attractive to women. Now I consider being short and heavy an advantage. I don't have anything to do with women. I hate their guts. I get such a revulsion when I see how women act that I can't bring myself to go after sex." (Quotations taken from Stunkard and Mendelson, 1967.)

You can see that overweight is far more than the comical cosmetic problem many people take it to be.

Besides the psychological and social problems related to obesity, there are serious problems of health. Obesity encourages a variety of serious diseases, including diabetes, high blood pressure, and heart disease. Given the seriousness of the implications of obesity, it is not surprising that experimental psychologists have turned their attention to studying its causes and potential cures.

IS OBESITY A PSYCHOLOGICAL PROBLEM?

The attitude generally taken in our society is that obesity is a moral problem. This is one of the reasons we tend to look down on the obese. "It's their own fault," we say. "They're too fat because they stuff themselves and have no will-power." In contrast to this, the scientific attitude is to look for the controlling variables responsible for making a person fat. The better we understand the variables that determine whether a person will be fat or not, the more likely we are to be able to control weight.

Until fairly recently, the outlook for the obese was bleak. Stunkard (1972) put it succinctly: "Most obese persons will not stay in treatment for obesity. Of those who stay in treatment, most will not lose weight, and of those who do lose weight, most will regain it." Despite the constant barrage of fad diets giving rise to teeming businesses dedicated to publishing the latest methods, supplying the magical vitamins and food supplements, or providing special low-calorie foods, overweight people seldom succeed in losing weight permanently. The outlook has now improved, largely due to the efforts of experimental psychologists (Stunkard, 1972).

Sometimes overweight is due to physiological abnormalities. Hormones play a role in determining whether food intake will turn to fat or not. Hormonal disorders may make it very difficult for certain people to lose weight. Mayer (1968) even describes an enzymatic defect that causes fat, once broken down, to redeposit as more fat. Mice with such a defect have been brought close to starvation while still retaining an abnormal amount of bodily fat. Physiological disorders of these types are probably best treated medicinally rather than psychologically. But they are relatively rare disorders. Most overweight people have a behavioral problem.

This is not to say that the overweight have no physiological "excuse" for their excess pounds. We know that the number of fat cells (adipocytes) possessed by a given person is determined early in life. That number appears to be unchangeable. So a person who was overweight as a child will only be able to lose weight by shrinking those fat cells to an abnormally small size.

In contrast, the person who became overweight in adulthood has not too many, but abnormally large, fat cells, and can get back to normal weight simply by bringing those cells down to normal size. Thus, some overweight people, specifically those who were overweight as children, have a special bodily disadvantage when it comes to losing weight. But we know that behavioral and bodily states heavily influenced by physiological causes can often be modified through learning. And current evidence suggests that, despite the physiological disadvantages, people with obesity of early onset can lose weight and keep it off.

PECULIARITIES IN THE OBESE WITH RESPECT TO CONTROL OF FOOD INTAKE

There is a long tradition of belief that overweight people eat for different reasons than people of normal weight. It is said that they eat because of "oral fixation," when depressed, when bored, when anxious, and the like. Research may eventually confirm some of these notions; but for now, the evidence in their favor is not strong. Therapies based on such ideas have focused on helping overweight people gain insight into their postulated fundamental problem without doing much, if anything, to help them to control the eating itself. Such therapies have been notable primarily for their failure to achieve weight loss (Stunkard, 1972).

There is one study, very much out of the mainstream of psychological research on obesity, that should be mentioned. It indicates that there may be a relationship between anxiety and overeating. Bornstein and Siprelle (1973) measured the effect of "induced anxiety" on weight loss in the obese. The "induced-anxiety" technique devised by Siprelle (1967) is one that helps people to identify and feel states of anxiety, and to make responses of relaxation to them.

Bornstein and Siprelle (1973) randomly assigned subjects from stratified blocks (based on percentage overweight) to one of four groups. They were a control group placed on a "waiting list" for later treatment, a group given insight therapy, a group given training in relaxation (progressive relaxation; Jacobson, 1957), and a group given treatment by the induced-anxiety technique. The group given induced-anxiety training lost an average of about 12 pounds and maintained the loss through a six-month follow-up. No other group differed significantly from controls in weight loss. It is interesting that these subjects averaged over 50 percent excess bodyweight. Such superheavyweights tend to come from the class of persons who have been obese since childhood, and have excessive numbers of fat cells.

Most research on obesity has focused not on such internal dynamics, but on the type of stimulus control to which the obese are susceptible.

Schachter's hypothesis

A major event in the experimental psychology of obesity occurred when Stanley Schachter proposed that overweight people are characterized by a general tendency to be under the control of external stimuli rather than internal stimuli.

Schachter's argument is that some people do not learn how to label the internal physiological state of hunger appropriately. They might confuse hunger with any state of arousal. It is not hard to see how such inappropriate attitudes could be learned. One possible source of it might be the practice of some parents who feed an infant that shows any sign of distress.

Some evidence in favor of the view that overweight subjects do not know how to label internal "hunger" states came from the work of Stunkard (1959; 1961), who measured gastric motility directly and related its occurrence to verbal reports of hunger. Subjects came to the laboratory in the morning without breakfast, swallowed a gastric balloon, and were asked every 15 min whether they were hungry. Thus, continuous records of gastric motility could be compared to the verbal reports of hunger. Overweight subjects differed markedly from normal subjects in the degree to which hunger reports corresponded to gastric movements, showing far less correspondence.

Schachter (1967) describes experiments done in his laboratory that permitted assessment of the effects of stomach loading and fear on eating in normal and obese people. Subjects were asked to skip a meal before coming into the experimental setting. They were told that the experiment was on the interaction of taste with other sensory experiences. Subjects in the "full" condition were told that they would be fed in order to "guarantee that your recent taste experiences are entirely similar." They were fed all the roast beef sandwiches they wanted and asked to fill out a long food-preference questionnaire. In the "empty-stomach" condition, they simply spent the period filling out the questionnaire.

The "fear" condition was manipulated by informing the subjects that they would receive electric shocks because electricity was being used as the tactile stimulus whose interaction with taste was of interest. "Low-fear" subjects were told that only the very lowest levels of shock were needed to get the effect of interest and that they might at most feel a slight tingle. "High-fear" subjects were shown an ominous shock machine, 8 ft high, and told that the experiment required the use of high voltages. They were asked whether they had a heart condition, thus suggesting to them that a severe stress was impending.

Eating behavior was measured by presenting subjects with several bowls of supposedly "low-calorie" crackers and asking that subjects taste and rate them on a number of characteristics. Subjects were told to base their judgments on the eating of as many or as few crackers as

they wished. The number of crackers eaten was taken as the eating measure.

Normal subjects responded to stomach preloading and high fear by reducing their eating in the test situation. Obese subjects ate, if anything, slightly more in response to these two conditions.

Hashim and van Itallie (1965) report clear evidence that overweight human subjects behave in the same way as brain-damaged rats in eating only "tasty" foods. Normal and obese subjects were brought into a hospital setting in which their food intake could be observed and controlled. Baseline measures were taken with a regular hospital diet of modest caloric value; then subjects were shifted to a bland liquid diet. While on the liquid diet, they could eat as much as they wished. Normal subjects dropped from their normal caloric intake for a few days, but then returned to a fairly steady maintenance diet. The intake of obese subjects dropped precipitously to a very low level and stayed that way for the duration of the experiment. Thus, obese subjects, though they overeat when palatable foods are available, not only do not overeat but will eat far less than a normal subject when the diet is dull and uninteresting.

In order to evaluate further the stimulus control of eating, Schachter used the ingenious tactic of speeding up a clock so that it appeared to be past mealtime. Subjects were brought into an experimental setting and wired to an electronic gadget that was supposed to measure such things as heart rate and galvanic skin response. On the pretext of gathering baseline data, the experimenter left the subjects alone for an actual half-hour. In one condition the clock had been speeded so that it read 6:05; in the other condition it read 5:20 at the end of the "baseline" session. The experimenter then returned, munching crackers from a box and telling the subject that he should feel free to help himself. He gave the subjects a variety of personality tests and then left them alone with the crackers for another actual 10 min. The dependent variable was simply the difference between the weight of the cracker box before and after the 10 min.

Since normal subjects are heavily controlled by internal conditions and obese subjects are less so, it was predicted that these two types of subjects would differ with respect to the control exerted over them by the clock. In fact, the obese subjects ate almost twice as many crackers when they thought it was 6:05 as they did when they thought it was 5:20. Normal subjects actually exhibited the reverse tendency, apparently, as some of them said, because their dinner would be spoiled by such munching.

Fat rats and fat people

We have known for a very long time that damage at the base of the brain can produce extremes of obesity along with certain other symp-

toms. There is a medical disorder called Fröhlich's syndrome. It includes obesity and shriveling of the gonads. Such problems used to be attributed to damage to the pituitary gland (which lies above the roof of the mouth, intertwined with a part of the brain called the hypothalamus). Now we know that damage to the hypothalamus produces obesity. Many studies have been done on rats with damage to the "ventromedial" part of the hypothalamus and its resulting obesity.

Stanley Schachter (1971) pointed out a striking number of similarities between rats with lesions of the ventromedial hypothalamus and obese humans. Some of the similarities seem obvious. Both eat more of good-tasting food. Both eat more food per "meal." (A meal for a rat is defined as a "burst" of eating.) Both are less active, and both eat faster. But it is not so obvious that both will eat *less* bad-tasting food than normals. If you dilute food with something that impairs its tastiness, overweight humans, even as infants (Nisbett & Gurwitz, 1970), will share with the brain-damaged rats a tendency to reject it.

Many studies also show that overweight rats and people will tend to put out less effort for food than normals. Rats with ventromedial lesions will tolerate high fixed-ratio schedules (see Chapter 6) less than will normal rats. Obese humans eat more shelled almonds than normals, but eat fewer than normals if the situation requires that they shell their own almonds.

Both "ventromedial" rats and obese humans are more responsive to the threat of aversive stimulation than normal. And they both regulate subsequent caloric intake better in response to liquid than to solid food preloading. Actually Schachter (1971) identified a dozen or so similarities between obese humans and rats with lesions of the ventromedial hypothalamus. He could find no studies indicating discrepancies in their responses.

We should not jump too hastily to the conclusion that obese humans suffer from a fault of the hypothalamus. Schachter's generalizations are fascinating and important. But they often rest on analogy. And there is some puzzling evidence of limitations on the generality of some of the findings mentioned above. For example, Cruce et al. (1971) have demonstrated that genetically obese rats of the Zucker strain work well on high fixed ratios and also adjust their food intake for calories. Why do genetically obese rats differ from both lesioned rats and obese humans?

There is also some evidence that the overeating and overweight following damage to the ventromedial hypothalamus occur fully only in adult female rats (Kurtz et al., 1972). The effects seem to be blocked by the presence of growth hormone. Male rats do not stop growing as they mature, but females do. Hence, neither sex shows the effect in infancy, and males never show it.

Schachter's generalization remains a fruitful basis for new experimental undertakings, but cannot yet be taken as hard fact.

Experimental attempts to evaluate Schachter's hypothesis

Schachter based his notion that the obese are abnormally prone to come under the control of external stimuli on a good deal of evidence. Much of that evidence has already been described, and it seems strong.

The tendency to external control goes beyond situations involving eating. Pliner (1973) did an interesting study in which she assessed the ability of the obese versus people of normal weight to think about an assigned topic. She assessed the relative success at sustaining such thinking by a method of converging operations. Subjects gave verbal reports of the amount of time spent thinking about the assigned topic, and their ability to distract themselves from a painful cold pressor stimulus by thinking of the assigned topic was also measured in a variety of ways.

There were three main experimental conditions given to obese and normal subjects. In one condition, there was an assigned topic and there were external cues to aid thinking about the topic. In a second condition there was an assigned topic, but no external cues were provided. In a third condition there was no assigned topic. With cues absent, obese subjects were much worse than normals in keeping to the assigned topic. With cues present, they were better than normals.

Schachter's hypothesis does not just say that obese people are more strongly controlled by external stimuli. It also says that people of normal weight regulate their eating according to internal caloric need, whereas obese subjects are less adept at this. Experimental evidence has accumulated to indicate that neither the obese nor those of normal weight are really very good at detecting caloric values on the basis of internal cues alone (see Wooley, 1972; Wooley, Wooley, & Dunham, 1972a and 1972b).

The striking incapacity to regulate food intake according to calories has been demonstrated by controlling for sensory cues normally associated with caloric value. Solutions of markedly different caloric value were devised so that subjects could not tell which of them had the higher caloric value. Subjects' beliefs about the caloric value of the solutions proved more closely related to their subsequent hunger than did the actual caloric values. This was true whether they were obese or not.

Devendra Singh has argued that the hypothesis of external control must at least be supplemented by the notion that the obese have an abnormally hard time changing previously established habits. He and his associates have accumulated a good deal of evidence for this view. For example, Singh and Sikes (1974) examined the basis for the reluctance of the obese to put out effort for food. Schachter would presumably interpret this as being due to the remoteness of food cues when, for example, almonds must be shelled. Singh and Sikes showed that such responses depend on what subjects are accustomed to. They

Key Ideas Box 13.1: Obesity as a psychological problem

Obesity is related to social, psychological, and physical problems to a degree far greater than most people realize. To some extent it is a physiological problem. For example, those who have been overweight since childhood have an excessive number of "fat cells" that cannot be eliminated. Reduced activity and improved maintenance of bodily temperature, which are concomitants of excess weight, tend to keep the weight high. For a person with obesity of early onset to maintain normal weight, it may be necessary to remain in a state of physiological starvation.

Schachter has pointed out numerous resemblances in the behaviors of rats that are obese as a result of damage to the ventromedial region of the hypothalamus and obese humans. For example, they eat more rapidly, take fewer meals, and are more finicky about what they will eat and less willing to work for it. Schachter argues that they appear to be abnormally controlled by external stimuli (including such things as taste) rather than internal bodily cues.

Research subsequent to Schachter's hypothesis has sometimes supported it, but sometimes not. For example, there is evidence that neither the obese nor those of normal weight can regulate caloric intake by internal (caloric) cues alone. Certain genetically obese rats fail to show the patterns described by Schachter. If trained to put out large effort for food before surgery, rats with ventromedial lesions choose a more in preference to a less effortful response to get food. There is evidence that they and obese humans have a problem of perseverating in established habits, even when maladaptive.

More research is needed to learn the exact role of external stimulus control in overeating.

observed that obese subjects were as likely as those of normal weight to unwrap chocolates, but much less prone to unwrap foil from cashews. They argued that this was due to the familiarity of foil wrapping on chocolates. However, more evidence is needed to make this conclusion firm. The greater incentive value of chocolate needs to be controlled.

Similar results have been obtained with ventromedial rats. If trained on tasks of high effort *prior* to surgery, they actually choose a more effortful response over a less effortful one in a free-choice situation (Singh, 1972).

Obese humans have difficulty changing from pre-established habits, even when the tasks are unrelated to eating (Singh et al., 1973).

Methodology and the study of obesity

There are numerous methodological problems in the study of obesity. *Measurement of obesity* is a problem. Standard weight charts are notoriously inaccurate, failing to differentiate the type of body tissue giving rise to the weight. Accurate methods are available, but logistically difficult. No generally established criterion has been accepted by researchers.

Biased samples of subjects are commonly used. For example, women and the relatively affluent are used out of proportion to their representation in the population of the obese. The effects of preloading of food have been studied without *controlling for composition* of the food or *correcting for differential caloric requirements* of obese and normal subjects. Measures are often taken in the *morning*—a time when the obese are less prone to eat. Little effort has been made to control for *demand characteristics.* Virtually all studies have been of *short-term* regulation of intake, whereas caloric adjustments are probably long-term in reality.

Control subjects are commonly put on a *"waiting list."* They commonly gain weight while on the list. This may be due to letting go of control in anticipation of the forthcoming program. The result is that experimental subjects are compared with controls who are artificially gaining.

Obedience to repugnant commands

The attention of the world was focused on the nature and causes of obedience when the Nazi party came to power in Germany and began implementing the "final solution to the Jewish problem." Millions of Jews were sent to their deaths, yet few German non-Jews seemed to be stirred by the atrocities enough to rebel or refuse to cooperate with them. After the war, soldiers were tried for war crimes generally having to do with the execution of noncombatants. In their defense, it was common for them to claim that they were simply taking orders and that responsibility for their acts rested with those who initiated the orders.

But why did they obey? Perhaps they feared for their own lives. It is likely that refusal to obey orders would have been followed by severe sanctions. Yet there were many who found ways to avoid participating in the atrocities and who nevertheless escaped severe penalties. Some people will argue that it had to do with the German "national character." To be sure, Germany before the Nazis was a highly civilized nation—a nation capable of producing a Bach and a Beethoven, a Goethe and a Rilke. But there seems to be a strange mixture of hardness and sentimentality in the Germans. It is difficult for one

raised in an Anglo-Saxon culture to keep from blushing at such German literary works as Goethe's *Sorrows of the Young Werther.* This sentimentality seems to be one side of the German character. But on the other side is a disciplinarianism and even a militarism that have few parallels in the Western world.

Is it the case that the dreadful acts of obedience performed during that war were the product of the strangely split national personality of the Germans? It has long been believed that "it could only have happened in Germany," and that the easy-going Americans, with their democratic traditions and inculcation into the principles underlying human dignity, would be incapable of obeying an order to slaughter noncombatants! In the past we would have made such a statement with great confidence. But today we must do more than hesitate over accepting it. Apparently American soldiers in Vietnam did in fact obey orders to kill noncombatants, including small children.

Anyone who has read Stanley Milgram's reports of his research on obedience and action conformity (the first of which was published in 1963) would not be terribly surprised to learn that Americans would obey orders to perform acts that are morally repugnant to them. His studies showed a remarkable, even astonishing, willingness to obey such commands. Milgram's experiments show how an ingenious experimenter can find a way to study even those phenomena that seem least amenable to investigation in the experimental laboratory.

The situation used by Milgram in these investigations was one in which subjects were required to deliver what appeared to be severely punishing, even dangerous, shocks to a "victim." In his first experiment, Milgram (1963) obtained subjects from the New Haven area through a newspaper advertisement or by direct solicitation through the mail. Those who responded to the appeal were led to believe that they were to participate in a study of learning and memory conducted at Yale University. They were paid $4.50 for participating, but were informed that the money was given to compensate them for coming to the experiment and that they could keep it no matter what happened after they arrived. Thus, even the modest financial reward was not contingent on their performance in the laboratory. Subjects came from various walks of life; 37.5 percent were "blue collar" workers, 40.0 percent were "white collar" workers, and 22.5 percent were professionals. Subjects were told that the experiment required that electric shocks be delivered to a learner whenever he made an incorrect response on a verbal learning task. The shock apparatus was shown to them, and it was labeled in such a way as to indicate that its range extended into highly dangerous shocks. There were 30 switches ranging from 15 volts to 450 volts. Along with the voltage indications, there were more descriptive labels, ranging from "slight shock"

through "moderate shock" to "Danger: severe shock," and finally two switches merely labeled "XXX."

Each subject was introduced to another "subject" who was in fact a confederate of the experimenter. He was an amiable fellow, generally regarded as a likeable sort. The confederate and the real subject drew lots to determine who in the experiment would act as "learner" and who would act as "teacher." The draw was rigged so that the real subject always acted as teacher. Following the assignment of roles in the experiment, the learner was strapped into an "electric chair," and the "learning experiment" began.

During the experiment the confederate arranged to make incorrect responses on three out of four trials. For each error, the real subject was instructed by the experimenter to give him a shock by closing one of the 30 switches. After the first error, he was to close Switch 1 (15 V); after the second error, he was to close Switch 2 (30 V), and so on to the end of the experiment. If a subject hesitated to continue, the experimenter had a fixed repertoire of prods that he presented. The prods ranged from a simple "Please go on" to "You have no other choice, you *must* go on."

In the earliest reported experiment, the "victim" was remote from the "teacher"—in another room. The victim did not protest until a shock level of 300 volts was reached (he was, of course, not really receiving the shocks). At that level, he pounded on the wall of the room in which he was strapped to the electric chair. The pounding could be heard by the real subject. After pounding, the learner's responses no longer appeared on the panel designed to indicate them to the teacher. The experimenter thereupon instructed the teacher to count failures to respond as errors and deliver the shock anyway. At 315 volts the victim pounded once again, but made no response thereafter. It is perhaps worth mentioning that the authenticity of the shocks was emphasized to the real subject as a preliminary to the experiment. Subjects received a true 45-volt shock as a sample before beginning.

The main dependent variable was the number or percentage of subjects who refused to go on after each of the 30 shock levels. Of the 40 subjects, 26 obeyed all the way to the end of the shock scale. This occurred despite the fact that there was good evidence that subjects believed that the situation was an authentic one. Subjects were asked after the experiment to rate the painfulness of the last few shocks they had delivered. The modal response was "extremely painful." Furthermore, subjects showed signs of extreme tension during the experiment. Many subjects were seen to sweat, groan, bite their lips, and engage in nervous and sometimes hysterical laughter.

Milgram (1965a) reports an interesting comparison between the frequencies of willingness to obey experimenter commands predicted by 40 psychiatrists and the actual frequencies. If anyone should know

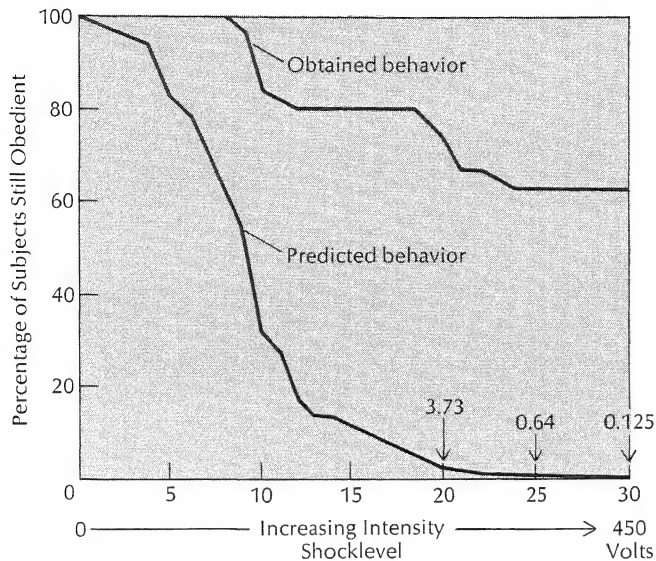



FIGURE 13.1 Comparison of degree of obedience predicted by a group of psychiatrists with degree of obedience actually obtained in Milgram's experiments. (From Milgram, 1964.)

about the skeletons in our intrapsychic closets, it should be men who listen to our most hidden secrets in psychotherapy sessions. Yet Figure 13.1, which compares the predicted with actual behaviors, indicates that even the psychiatrists very greatly underestimated the willingness of subjects to obey such commands. For example, they estimated that approximately 1 of 100 people would go to the end of the scale, whereas over 50 percent actually did so.

THE ROLE OF INSTITUTIONAL PRESTIGE

Milgram's original experiments were conducted at Yale—admittedly a highly prestigious institution. It would be reasonable to suspect that subjects obeyed instructions because they could not believe that an experimenter working under the auspices of Yale would do anything truly harmful to his subjects. In the introductory phases of the experiment, the experimenter said that no permanent damage would be done. How likely is it that a Yale experimenter would lie or be in error with regard to a matter of such importance? To some degree, this interpretation seems inconsistent with the indications of strong misgivings on the part of the subjects, but Milgram, not content with such inferential evidence, did an experiment to assess the role of institutional prestige in obtaining obedience.

The original experiment was replicated under conditions that gave no indication of any connection with Yale. A modest laboratory was set up in rented quarters in Bridgeport. Subjects were obtained through a mail circular sent out by "Research Associates of Bridgeport," which was an organization concocted for the experiment. Other than the change in setting and affiliation, the experiment was a close replication of the earlier study. Surprisingly enough, there was no significant reduction in obedience in this new setting. Apparently the readiness with which such orders will be obeyed is even greater than might be inferred from the earlier study.

VARIABLES THAT INFLUENCE OBEDIENCE

Closeness of the victim

Though prestige of the institution proved to be ineffective as a determinant of obedience, certain other factors did have an influence. One of these was the degree to which the teacher had contact with the victim. Four degrees of proximity were used:

1. A remote condition (the one already described) in which the victim merely pounded on the wall of his room.
2. A voice-feedback condition in which the learner's complaints were heard. These complaints included proclamations of refusal to continue in the experiment, cries of "I can't stand the pain" and agonized shrieks.
3. A physical-proximity condition in which the victim was in the same room with the teacher, only $1^1/_2$ ft away. In this condition he also gave the voice cues.
4. A touch-combined-with-proximity condition, in which the situation was that of the proximity condition but in which it was also necessary for the teacher to hold the victim's hand down on the shock plate.

The findings were that the mean maximum shock delivered decreased with increases in teacher-victim contact. Figure 13.2 illustrates this relationship. The percentage of obedient subjects for the four conditions was 66 percent for the remote condition, 62.5 percent for the voice-feedback condition, 40 percent for the simple proximity condition, and 30 percent for the condition that combined touch with proximity.

Closeness of authority

A second influential variable was the degree of contact that subjects had with the experimenter. Only two levels of this variable were used. In the first, the experimenter was present while giving his instructions.

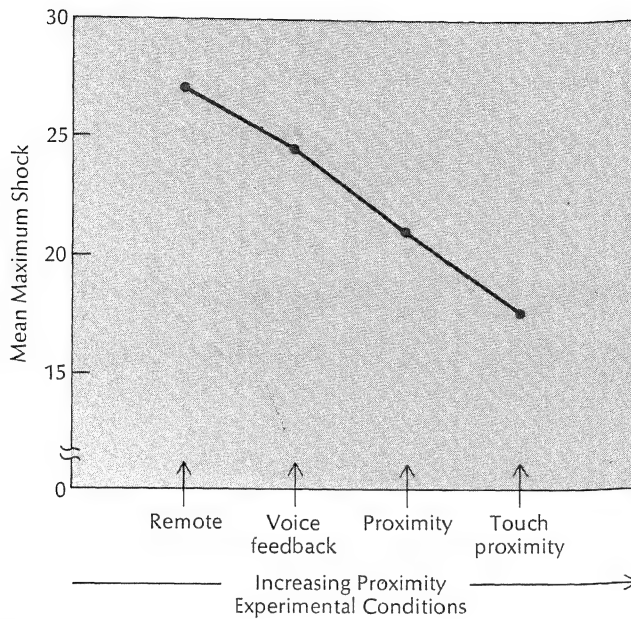



FIGURE 13.2 Mean maximum shocks delivered by subjects to victims, varying as a function of the "nearness" of the victim. (From Milgram, 1965b.)

In the second, the experimenter left after introducing the subject to the confederate, and all further communications were carried out over a telephone. When instructions were given over the phone, there was only about one-third the obedience found when the experimenter was physically present during presentation of the orders.

An interesting behavior was observed when instructions were given over the telephone. Some subjects told the experimenter they were complying with his instructions, but actually held the shock levels down. Apparently these subjects were more concerned with appeasing the experimenter than with seeing to it that the purposes of a scientific experiment were fulfilled.

REPLICATIONS OF MILGRAM'S FINDINGS

Milgram's findings have been replicated a number of times, both by Milgram himself and by others. These include studies by Haas (1969), Tilker (1970), and Holland (1967). In general, where groups truly comparable to those run by Milgram have been tried, the results have been very much like those obtained by Milgram. In fact, levels of obedience have tended, if anything, to be somewhat higher in replications subsequent to the Milgram studies. King (1971) has provided an excellent summary of the studies under discussion here (Table 13.1).

TABLE 13.1 Obedience Levels Obtained by Several Experimenters Using the Milgram Paradigm

	Milgram (1963)	Milgram (1965b)	Haas (1969 and more recent)		Tilker (1970)	Holland (1967)
Remote	65%	66%	66%	100%	93%	
Voice		62.5%			73%	75%
Proximity		40%	26%	70%	47%	
Touch		30%				

Note: Type of feedback from the victim is indicated at the left of the table.
From King, 1971.

The use of role playing in studies of obedience

At least two studies have used role playing instead of or in addition to deception in order to determine levels of obedience to malevolent authority. Holland (1967) had three groups: one a direct replication of Milgram's auditory feedback condition, one in which subjects were told that they were to figure out what the experiment was really all about and thereby find out how a real subject would respond, and a third in which subjects were told that actual shock levels were only 10 percent of what the labels said, but that they were to play the role of a regular, naive subject. Holland concluded, on the basis of his results, that none of these conditions made any statistically significant difference.

Influenced by these findings, Orne and Holland (1968) went on to argue that Milgram's findings were invalid because they were due to subjects' response to what they construed to be the experimenter's hypothesis. Duplication of the Milgram findings by role-playing subjects does seem to lend credence to their argument.

However, King (1971) reanalyzed the data from Holland's (1967) study, pointing out that Holland had compared his own voice-feedback conditions to data taken from Milgram's no-feedback condition. Comparing appropriate conditions, King found that differences emerged. For example, Holland's findings for the group told that shock levels were at 10 percent of the labeled values differed from the comparable condition in Milgram's studies, with $p < 0.01$. The analysis by King suggests that the Milgram findings are *not* duplicated by subjects attempting to play the role of naive subjects, and weakens the argument made by Orne and Holland (1968) that the original findings were due to subjects responding to demand characteristics.

A second study using the role-playing technique was done by O'Leary, Willis, and Tomich (1970). Essentially, their experiment was like that of Milgram (1963) except that the subjects were permitted to *read the method section* of the experimental proposal prior to participating in the experiment. Furthermore, they were allowed to ask questions about the method, and were given honest answers to the questions. They were simply instructed to "role play"—to behave as

they would if they believed that the experiment really involved delivery of shocks. The results were virtually identical to those of Milgram (1963). Subjects also showed many of the other behaviors described by Milgram, such as tension and laughter.

Unfortunately, it is difficult to interpret these findings. Willis, O'Leary, and Tomich (1970), suggest that the subjects may not have believed that the shocks were not real, even though this was stated in the material they read. Subjects expect to be deceived, and most of them probably know how unusual it is for an experimental psychologist to allow the subject to know the study's "gimmick." Thus, they may have perceived the study as involving a very real and dangerous situation. King (1971) argued that the Holland (1967) study provided a much more plausible rationale, telling subjects that they "had to figure out what the experiment was all about" and that the shocks were only 10 percent of the labeled value. But this is speculation, and the argument against the ecological validity of Milgram's studies must be taken seriously.

Are the Milgram studies ecologically valid?

Orne and Holland (1968) have argued that Milgram's findings reflect subjects' responses to the demand characteristics of his experimental situation rather than the realities of obedience. They maintained that the situation was implausible, since the experimenter required nothing of the subject that he could not easily do himself and the experimenter's willingness to tolerate the victim's suffering was given no justification by emphasizing the "world-shaking importance of the learning experiment." Orne and Holland state: ". . . that the *S* will in an experiment carry out behaviors that appear destructive either to himself or others reflects more upon his willingness to trust the *E* and the experimental context than on what he would do outside of the experimental situation" (p. 291).

Support for their argument comes from the observations that Milgram's findings can be duplicated, or virtually so, by role-playing subjects (Holland, 1967; Willis, O'Leary, & Tomich, 1970). However, the clarity of this support is marred because Holland's results were significantly different from Milgram's in at least some cases, and because the subjects of Willis, O'Leary, and Tomich may have believed they were being deceived (King, 1971).

Persuaded by the arguments of Orne and Holland (1968), Sheridan and King (1972) decided to conduct a systematic replication of Milgram's work, substituting an *authentic victim* for the usual confederate. The victim was a cute, fluffy puppy. Their reasoning was essentially that subjects could not dismiss the pain of the puppy by believing that it was acting, and they would therefore refuse to carry out the shock procedure.

A plausible reason for having subjects throw the switches was provided by explaining the dangers of experimenter bias to them, showing them a discussion of such bias taken from Sheridan (1971). The puppy was supposedly learning a discrimination of the rate of flicker of lights. In reality, it was impossible to solve the problem because there was no difference in rate of flicker. The subject could not judge this, however, because of the dim lighting and because the puppy was viewed through a double layer of glass.

In order to protect the puppy from harm, the shock values were not really those indicated on the switch panel. Actually, they were of very high voltage but very low current. This kind of shock stings but does, at most, very minor tissue damage. Nevertheless, the puppy became agitated, barked, and even howled in response to the shocks. The amount of shock really did increase as the higher switches were thrown, and the puppy's responses became more intense.

Both male and female college students served as subjects, and, much to the surprise of Sheridan and King, typically did not refuse to shock the puppy. Male subjects closely replicated the comparable results obtained by Milgram with a human confederate as victim, and 100 percent of the women shocked the puppy all the way to the end!

Sheridan and King had a variety of ways of detecting whether subjects believed in the situation, including postexperimental questioning ("What did you think the *real* purpose of this experiment was?"), measurement of how long subjects held the shock switch in place (they tended to try to minimize the duration of shock more and more as the levels grew higher), and a separate interview by a teacher lecturing on the Milgram studies who appeared unrelated to the experiment. It seems that the authenticity of the experiment was quite convincing.

Perhaps the best index of ecological validity is the ability of an experimental model to predict behavior in real situations. Orne and Holland (1968) argued that the behavior of subjects in Milgram's studies reflected more on their willingness to trust the experimenter than on what they would do *outside* the experimental situation. As King (1971) pointed out, ". . . the tragedy of My Lai was taking place at this time 'outside of the experimental situation.'"

Are women more obedient than men?

Sheridan and King (1972) got significantly more obedience from female than male subjects. Every single woman complied to the end of the scale. Results of at least one other study were virtually identical to this, with only one woman refusing to go on (Goldman, personal communication). However, a number of other studies have not indicated greater compliance on the part of women (see, for example, Killham & Mann, 1974; Milgram, 1974). We do not know why the results differ

Key Ideas Box 13.3: Obedience to repugnant commands

Stanley Milgram devised an ingenious tactic for studying willingness of subjects to obey repugnant commands. He required them, as part of an experiment, to *deliver increasingly severe shocks* to a "learner-victim" (in fact a collaborator with the experimenter). The shocks were purportedly dangerously severe, but a majority of subjects obeyed. Milgram varied such things as the prestige of the research establishment, the proximity of the victim, and the proximity of the experimenter. Some of these had effects, but considerable obedience still occurred. His findings ran contrary to expectations of behavioral experts, which he measured.

Objections to the validity of Milgram's studies have centered on the possibility that demand characteristics influenced responses. Studies in which subjects were merely asked to role play or were told that the shocks were only 10 percent of the indicated value produced much the same results as Milgram had gotten. The ecological validity of Milgram's findings is not easy to establish, but one study produced very similar results with an *authentic victim* (a puppy).

A method for studying obedience to commands in *natural settings* has been devised. People are simply told to do such things as pick up litter or give someone else a dime for a phone. Subjects are notably compliant, especially when the person giving the commands wears a uniform.

across experiments, though it may be due to differences in age. The highly compliant women were quite a bit younger than the less compliant ones.

A METHOD FOR STUDYING OBEDIENCE TO REPUGNANT COMMANDS IN THE FIELD

Bickman (1974), who was primarily interested in the influence of type of clothing as a social stimulus, devised a method for studying obedience in a natural setting. The experimenter merely walked up to people on the street and directed them to perform some mildly distasteful act. They were told to pick up a paper bag littering the street, to give someone else a dime for a pay phone, or to stand in a different position with respect to a street sign (obviously intended for automobiles, not pedestrians) that said "No Standing." Though Bick-

man's interest lay in the influence of a uniform on obedience to these commands, the method is quite generally applicable.

Bickman found that a person wearing a policeman's or guard's uniform was far more likely to be obeyed than someone in regular clothing. This occurred even though the commands clearly exceeded any real authority inherent in the role identified by the uniform. The compliance of subjects in this situation occurred even when the uniformed agent abruptly left after giving the command. Hence, it appears not to be due to some fear of reprisal.

Study Questions

1. To what extent is obesity a physiological and to what extent a behavioral problem? Give evidence.
2. Explain Schachter's view of peculiarities of stimulus control in the obese and cite the research literature on which he based that view.
3. Summarize the experimental evidence used to evaluate Schachter's hypothesis since it was put forth.
4. Summarize the main research done on methods of controlling overeating based on experimental psychology.
5. Outline the major methodological problems that must be confronted in doing research on methods of controlling weight and on functional differences between the obese and those of normal weight.
6. Explain Milgram's method of studying obedience and give his main findings.
7. What variables influence obedience?
8. What results are obtained when role playing is substituted for deception within the framework of Milgram's tactic?
9. What methodological objections have been made to Milgram's studies?
10. Are Milgram's studies ecologically valid?
11. What do we know about the influence of sex on conformity and obedience?

14 GETTING IDEAS FOR RESEARCH

You may attend class conscientiously and learn a great deal about experimental psychology, then draw a blank when asked to come up with an experiment of your own. There are many obstacles to overcome in getting good research ideas. Here I will give some suggestions.

Ideas come from reading scientific literature

If you read scientific books and articles with an active interest, many prospective research problems will be likely to come to mind. This happens even to students reading their first psychology text. However, it is very important to keep a *log of research ideas.* This will serve to fix the ideas in your memory, and it will pull your ideas together in one place. Often pieces will fall together in this way, and sometimes you will notice the same idea recurring in various forms. Furthermore, even if we were once loaded with ideas for research we may later find ourselves at a loss when we have the opportunity or the obligation to do some research. The log will remind you of the ideas you once had.

It is hard sometimes to read research literature until you have done some research yourself. A typical pattern for my students has been to begin a required research project reluctantly, then to be sparked by an exciting new idea from related literature they might otherwise have found tedious.

SEARCHING THE LITERATURE

The following information will help you to familiarize yourself with the literature in a given area of interest.

The first and most general sources will be *textbooks,* such as this one. These will give you a general notion of the kinds of questions that have interested research psychologists, and may stimulate you to extrapolate from previous research.

Once you have a basic notion of the area in which you would like to read, take advantage of one of the most valuable resources in literature searching—*well-informed people in the field.* Students often neglect to make use of such expert sources as their teachers or other faculty members at their college or university. These experts probably have deep knowledge of the area of interest, and are likely to be very well aware of available literature sources.

You and the experts may be able to become more specific about the topic of interest. You can make explicit the various forms in which it might be referenced. The more explicit you can get, the easier it will be to identify sources pertaining to your topic.

Look in *library card catalogues* under the topic areas of your interest. Books are exceptionally good starting sources, since they will provide large numbers of references to relevant articles you can look up.

Psychological Abstracts is a journal available at most college and university libraries that indexes and briefly summarizes most articles in psychological journals. Look in the index of *Psychological Abstracts* under the various names you think might be used for research in your topic area. This will refer you to pages containing summaries of the articles. Having read these abstracts, you can determine whether to search out an article itself. Other reference works such as *Biology Abstracts* or *Index Medicus* may be useful for certain topic areas in psychology.

Review articles are published in several sources, notably *The Annual Review of Psychology, Psychological Review,* and *Psychological Bulletin.* Review articles are valuable in giving you a general perspective on a topic area, and in providing many further references. However, do not rely entirely on them, since an article may not be well represented by the author, who often will have a strong theoretical point of view to support.

It is extremely important that you seek out articles that are referenced in the articles you first read. You look up the references in a given article, then look up the references in those articles, then look up the references in those articles, and so on until you seem to have covered all the major articles. I have found this to be an extremely effective way of reviewing the literature, and would rely on it far more than the fallible abstracts or computer retrieval services.

You may have access to various computer retrieval services, such as Datrix (University Microfilms, Inc., Ann Arbor), which provides references to most Ph.D. thesis abstracts in North America, or to Medlar, which searches *Index Medicus* for the last two years. These systems rely on coding of articles with key words. The mindless computers, though capable of reviewing enormous quantities of information, may give you a lot of chaff with the wheat, and also may miss some wheat, so don't overvalue this impressive resource.

Several periodicals regularly publish lists of articles by topic area and/or author. Two important ones are *Current Contents—Life Sciences* (Institute for Scientific Information, 325 Chestnut Street, Philadelphia, Pa. 19106) and *Science Citation Index.* These will probably be in your library. Others may also be available. Check with your librarian.

It is a regular practice for authors to provide without charge copies of their articles (reprints) and even copies of articles they have done that are not yet printed (preprints). You can request these either by postcard or by letter. It is deferential to send a letter, so you should do that if you are making a rather large request. Some authors will not have reprints and preprints, especially in certain countries. But it doesn't hurt to ask.

If you have specific needs for information, you should not hestitate to get on the phone and call an author. In science, even a humble, beginning student can request information from a Nobel prizewinner with good results.

HOW TO READ ARTICLES IN EXPERIMENTAL PSYCHOLOGY

Most articles in experimental psychology are not easy to read unless you have a good deal of background. Costs of publication force editors to keep articles down to the bare bones. Keeping certain key questions in mind will help you. First, ask, "What is the problem to be solved, and why is it significant?" The article may not give you much help in answering this question, because authors are expected to write for an audience somewhat familiar with the topic of research. You may have to ask teachers or look for background reference materials before really grasping the significance of the paper. Try not to fall into the fallacy of judging that the article is "trivial" because you do not immediately see why it is significant. It sometimes boggles the mind to see how important a superficially trivial article can be once the background of the problem is understood. One of the advantages of your having had the kind of education you have had is that you can look beyond the superficialities.

Next, ask "What is the tactic, what is the experimental design, and

Verification and replication

The simplest extrapolation from previous research is to *verify* it. An experimenter who *replicates* a previous finding is verifying that finding. Replications make an important contribution to the advancement of scientific knowledge. Psychologists, especially journal editors, should encourage more of them. Many disciplines do not regard a finding as established until it has been "repeated in other laboratories." Unfortunately it is not easy to get simple replications published.

Variations on simple replication are possible, and of benefit. For example, it may be possible to repeat a finding and also to measure it in a slightly different way. Such studies are more likely to be published.

Recombination

When systematic replication is done, the experimenter has made a transition into a second mode of extrapolating from previous research, that of *recombining* aspects of previous research. With recombination you take parts of two or more previous experiments and combine them. For example, Donald Blough (1956) was interested in studying the sensory capacities of animals. But how do you ask animals what they see or what they hear? He was aware of the methods used by B. F. Skinner to study operant conditioning. Blough was also aware of certain methods devised by George von Bekesy to study sensory capacities of humans. Essentially, Bekesy had people indicate whether they could or could not detect a stimulus by holding down and letting up on a key. But the operant chamber typically requires animals to press or peck on a key. Why not apply the Bekesy method to animals in operant chambers? This is what Blough did.

Generalization

You can create hypotheses for research by generalizing from previous findings. Generalization can be across various aspects of an experiment. Will the same functional relationship hold with other species? Will it hold in other situations? Will it hold with other kinds of measures (such as a change from verbal to nonverbal measures or a change in type of apparatus)?

Specialization

Specialization is the opposite of generalization. How does a general phenomenon apply in special cases? For example, it is widely known that psychiatrists and psychologists have a higher rate of suicide than the average population. What happens if you look at men versus women? It turns out that the higher rates of suicide are entirely due to more suicides by women.

Unconfounding

A careful reading of the literature will expose many confounded variables, many of them deliberately so. The section on obesity illustrates this. More spectacular findings often have a greater-than-average tendency to have confounded variables, since they tend to be initial efforts toward working in a given area. Many good research projects can be created by putting in the needed controls.

Improving measurement of the variables

It is common for psychologists to rely on nominally scaled variables when variables scaled at a higher level are within reach. Or they might use a verbal or an obtrusive measure when better measurement is possible. Many interesting findings can come from improving measurements and then systematically replicating the earlier work. There is a particular need for studies with multiple, converging operations.

Identifying underlying causes

Research to identify the underlying causes of a phenomenon is probably the most advanced and most satisfying of all research. It commonly comes relatively late in a research attack. To illustrate, I did an experiment that showed that there is a marked difference between albino and pigmented rats in the ability to do visual discriminations with the part of the visual pathway that goes directly from eye to cerebral hemisphere without crossing to the opposite side (Sheridan, 1965). The first stage of follow-up research had to do with *verification* by replication. The study was repeated with slightly different rats, in a slightly different situation (Sheridan & Shrout, 1965). Next, the phenomenon was verified by establishing it with various measures of neuropsychological (Creel & Sheridan, 1966), anatomical (Lund, 1965), and electrophysiological (Creel, 1971) kinds. Next, but to a degree overlapping the verification stage, was a stage of *generalization.* Tests were done to see if the phenomenon applied to a range of other species. It is now known to apply to rats, guinea pigs, cats, minks, ferrets and humans.

But the thrust of contemporary research is toward identifying underlying *causes.* The first move in this directions came from Sanderson et al. (1972), who began detailed genetic manipulations. A variety of genetic experiments and biochemical experiments are now being done to find the underlying causes of the nervous-system peculiarities associated with albinism.

Ideas come from techniques and apparatus

Many experimenters center their research on a special technique or apparatus. Because of the success of the microelectrode technique,

many investigators think of themselves as "microelectrode people." Others are categorized as "operant conditioners" because their research is limited to the use of the operant apparatus.

There is an important distinction between *problem-oriented research* and *technique-oriented research.* In problem-oriented research the experimenter will use a variety of techniques, and even learn or devise new ones in order to get at the answer to a given question. Problem-oriented research might seem preferable, but it may not be easy in these complicated times to learn all the techniques required to solve a given problem. This leads to the notion that *research teams* of specialists would be ideal for solving research problems, but such teams rarely work out. I think this is because a really good researcher who is master of a given technique will not like to subordinate personal interests to goals imposed by others. First-rate scientists tend to be rugged individualists.

Though it might be reasonable to expect the inferiority of technique-oriented research, experience does not necessarily support that expectation. Some of the most important research done in recent times has come from technique-oriented researchers.

Ideas come from other people

People naturally want to do research on their own ideas, but there is nothing wrong with picking up a research problem suggested by someone else. Charles Martin Hall was a student in a chemistry class where the teacher suggested the importance of devising an economical way to process aluminum ore. Hall went right to work on the problem and licked it, making a great personal fortune. People often have good researchable ideas while lacking the will or even the capacity to implement them. Why not put these ideas to use?

Often you do not get people to just hand you a research idea, but you get ideas from them in less direct ways. Discussing issues, brainstorming together, and especially *arguing* about ideas are fertile sources.

With all due respect for your ideas, it is often good to postpone independent work, and to do research under an expert for a while. Most accomplished scientists have done so. Most people have to take a period of time learning a basic framework and developing skills and knowledge of techniques. If you try to go out too far on your own, you may get discouraged and lose interest in science. A lot of students feel guilty or thwarted or put down if they are denied an opportunity to conduct completely independent research. Actually, apprenticing yourself to an expert is just being smart and taking full advantage of the resources available.

RIGOROUS ESTABLISHMENT OF AN UNVERIFIED BELIEF

A fertile source of research ideas can be found in common beliefs that have not been experimentally established. Many such beliefs may be taken from *common lore.* Commonly held beliefs about animal intelligence led Edward Lee Thorndike to his research on animal learning. But *results of previous research* or *opinions of previous researchers* may also have been accepted more completely than the evidence merited.

It is widely accepted that the reason why rats learn visual pattern discriminations better in a Lashley Jumping Stand than in discrimination boxes is that the Jumping Stand requires that they "look before they leap." But this is only an opinion held by Lashley, and has never been established with experimental rigor. There are many such widely held opinions requiring experimental demonstration.

IDEAS COME FROM THEORIES

One of the great benefits of having a theory is its usefulness in providing many research ideas. Theories have implications, and experimenters can work these out and test them. Theories also have assumptions or postulates. In psychology, the assumptions or postulates have often been the focus of experimental testing. In the so-called "classical age" of learning theory, a great deal of attention was given to testing the postulates of psychological theories. This is an odd way to use a theory, and probably an improper one.

Many historically great theories have had false assumptions. For example, a look at Bernoulli's earliest formulation of the kinetic theory of gases reveals that every assumption ran counter to the facts. He assumed that all particles are perfectly elastic, that each particle has the

Key Ideas Box 14.1: Sources of ideas for research

The *scientific literature* is a major source of ideas for research. Changes in the type of design, adequacy of controls, measurement of the variables, methods of establishing reliability and generality, or in the power of the experiment provide new experiments. New *techniques* can be applied to problems old and new. *Other people* often have excellent ideas they do not want to implement or on which they want collaboration; conversation discloses many *unverified beliefs* in science and in the everyday world that can be put to the test. *Theories* can also be good sources of ideas.

same velocity as every other, that the particles never collide, and so on. Yet his was one of science's most successful theories, and an extremely useful one for predicting the behavior of gases. The implications held within a wide range of conditions (though, for instance, when pressure grows very great, the role of collisions of particles becomes too important to ignore).

So the heuristic use of theory (that is, its use as an aid to discovery) comes from the opportunities offered for testing implications. Controversy rages from time to time among psychologists on whether theories should be used to guide research, so let us go into the issue in some detail.

The role of theory in guiding research

CONVENTIONAL DESCRIPTION

There is a conventional description of science that says that experiments are done to test hypotheses and that hypotheses are derived by looking at the implications of a theory. This is illustrated in Figure 14.1. Hypotheses are simply "if–then" statements, such as "if I apply heat to a gas, then its pressure will increase"; "if light rays pass near

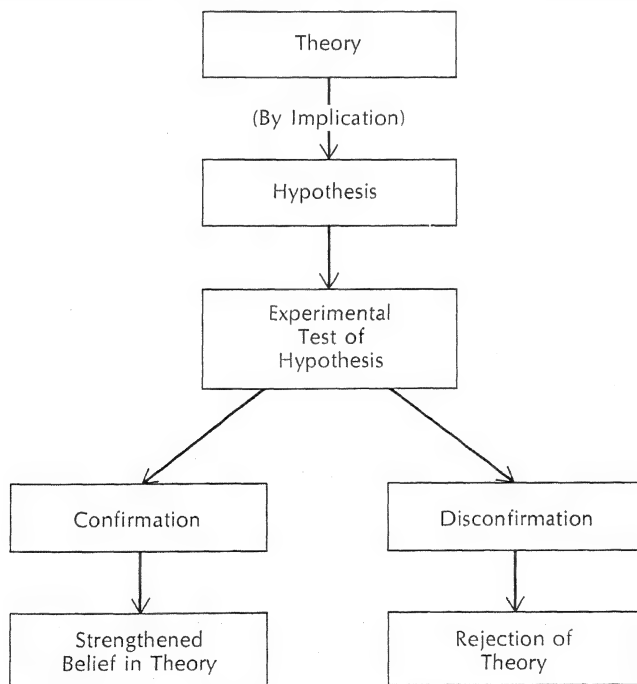



FIGURE 14.1 The "conventional description" of the scientific enterprise. See text for explanation and modern modifications of this view.

the sun, they will be deflected"; and "if people must overcome great obstacles in order to join a group, it will enhance the value of membership to them." If a hypothesis is verified, if observations correspond to it, the theory is *supported.*

The classic form of the conventional theory says that failure of experimental findings to fit the theory means that the theory is *refuted.* A confirmation only strengthens belief in the theory, without proving it correct. This is because many alternative theories might predict the same result. But a disconfirmation means that something necessarily implied by the theory runs counter to the facts. And it indicates that the theory cannot be correct. At a minimum, modification of the theory is required.

More recently philosophers of science have come to recognize that in fact scientists cling to a disconfirmed theory as long as there is no viable alternative theory. Theories are not rejected so much because of data inconsistent with them; they are rejected because a better alternative theory replaces them.

In psychology theories have declined in still another way. The learning theory of Hull seems not to have been clearly disconfirmed by data, nor to have been replaced by a better theory. Many findings seemed "kind of" inconsistent with it. Yet it could be stretched to fit most of them. The problem seems to have been that it did not provide a very useful model for predicting measurements not yet observed. The advantages of it were eventually dwarfed by the disadvantages inherent in being fettered by it. Most psychologists lost interest in it, and it went out "not with a bang, but a whimper."

Serendipity

Increasingly, scientists recognize how many discoveries have been made *serendipitously* rather than through theory. Serendipity is the art of finding something while you are looking for something else. Many, if not most, discoveries occur "accidentally," not by testing hypotheses born of theory. It was through serendipity that Pasteur discovered immunization.

> Pasteur's researchers on fowl cholera were interrupted by the vacation, and when he resumed he encountered an unexpected obstacle. Nearly all the cultures had become sterile. He attempted to revive them by sub-inoculation into broth and injection into fowls. Most of the subcultures failed to grow and the birds were not affected, so he was about to discard everything and start afresh when he had the inspiration of re-inoculating the same fowls with a fresh culture. His colleague Duclaux relates: "To the surprise of all, and perhaps even of Pasteur, who was not expecting such success, nearly all these fowls withstood the inoculation, although fresh fowls succumbed after the usual incubation period." This resulted in the recognition of the principle of immunization with attenuated pathogens (Beveridge, 1950).

Beveridge (1950) has compiled a sizeable list of cases of scientific discovery, and one cannot help but be impressed with the number of them that occurred serendipitously. Taylor et al. (1959) reported on a conference in which research workers in psychology described their approach to research; here, again, the description is one of unsystematic, intuitive groping toward an only vaguely defined goal. Young (1951) has given an excellent description of the research scientist:

> One of the characteristics of scientists and their work, curiously enough, is a confusion, almost a muddle. This may seem strange if you have come to think of science with a big S as being all clearness and light. There is indeed a most important sense in which science stands for law and certainty. Scientific laws are the basis of the staggering achievements of technology that have changed the Western world, making it, in spite of all its dangers, a more comfortable and a happier place. But if you talk to a scientist you may soon find that his ideas are not all well ordered. He loves discussion, but he does not think always with complete, consistent schemes, such as are used by philosophers, lawyers, or clergymen. Moreover, in his laboratory he does not spend much of his time thinking about scientific laws at all. He is busy with other things, trying to get some piece of apparatus to work, finding a way of measuring something more exactly or making dissections that will show the parts of an animal or plant more clearly. You may feel that he hardly knows himself what law he is trying to prove. He is continually observing, but his work is a feeling out into the dark, as it were. When pressed to say what he is doing he may present a picture of uncertainty or doubt, even of actual confusion (Young, 1951).

Theories may limit perspective

Theories have often been charged with limiting the experimenter's perspective. During the 1940s and 1950s, too much attention was concentrated on tests of theory. Other issues tended to fall by the wayside. And the value of the research done in response to theory rests heavily on the value of the theories, which is now in grave doubt.

Theories of color vision have dominated research on color. Researchers have emphasized testing of three-pigment versus opponent-colors theory. Other equally important questions have been given the once-over-lightly. What is the use of color vision? Why do people pay large sums of money for color television sets? Can color vision be modified by experience? These and many other questions have been researched much more lightly than the theoretical questions.

A similar tendency to put on blinders when in the framework of theory can be seen in many research areas.

THE USES OF THEORY

Modern experimental psychologists have, to say the least, trimmed theory down to size. The conventional description greatly exaggerated

the role of theories as heuristic devices. Scientists succeed in making discoveries without theories, and theories may even hinder them in some respects.

Granted that theories have been overblown as heuristic tools, should we conclude that the sole business of scientists is to gather the raw facts? May we conclude that theory making is useless? This is a conclusion that suits the temperament of many experimentalists but it is unjustified. Even if we grant that theories have *no* value as heuristic devices, they may well have other uses.

Theories can be mnemonic devices

A theory may be an aid to memory—a *mnemonic device.* Scientists complain of the enormous explosion of scientific literature, and all acknowledge that they cannot begin to read, still less to remember, a large number of otherwise diverse functional relationships.

Theories can predict limiting conditions

In a similar vein, a theory can provide rules for deciding when a given set of functional relationships will apply and when they will not. Functional relationships inevitably have *limiting conditions.* That is, relationships hold only if certain other variables are within reasonable limits. Unless we have a way to decide when they hold, the statement of functional relationships becomes trivial. Such statement boils down simply to "it holds when it holds and it doesn't hold when it doesn't hold." Then one can determine whether a given relationship holds only by observing what happens to the dependent variable when the independent variable has been manipulated. But this undermines the predictive value of the functional relationship. We are saying: "Nature will do whatever you see it do when you look." Reliance on raw experience without theoretical interpretation, when carried too far, becomes sterile.

Theories can make raw data usable

Theories can stipulate limiting conditions. For example, Bernoulli's kinetic theory of gases implies that the variable "frequency of molecular collisions" will exert critical influence when gas pressures are high, but not otherwise. It is said that, in the Middle Ages, if a person wanted to know how many teeth there were in a horse's mouth, he or she would look it up in Aristotle rather than take a look inside a horse's mouth. In reaction to this tradition, we now regard "the facts" as almost sacred. But it is important to realize that the value of facts lies in our ability to *use* them. One might well be better able to control and understand nature with fewer facts after the information has exceeded a certain limit. In Gogol's classic (and entertaining) novel *Dead Souls,* a strategy is described whereby guilty defendants could foil the courts through

flooding them with evidence. In the Russian courts of the time, a defendant could present as much evidence in his defense as he wished, and there was no penalty for presenting irrelevancies. Consequently, no one who could afford to pay the lawyer's fees need ever fear conviction because he could simply supply the court with reams of written materials that could not be read and evaluated in his lifetime. Theories, by simplifying large masses of data, can reduce the information that has to be processed, while retaining its pertinent aspects. Massive data gathering without theory may place twentieth-century science in the position of the courts of nineteenth-century Russia.

Theories can predict measurement in novel situations

Since theories can provide a rule for deciding when given functional relations apply, and because they have implications that are not always apparent from the simple original statement composing the theory, they can predict to novel situations. Predictions can be made on the basis of a theory, including predictions of what will occur in situations that have never been observed before. It is one thing to predict that what has happened previously under a given set of conditions will happen again if the same conditions are reinstated. It is another thing to predict what will occur in new, unprecedented situations, as theories commonly do.

Theories are sometimes valuable heuristic devices

It is not at all correct to suppose that theories are of no value as heuristic devices. There have been many cases in the history of science in which theories guided research. Mendeleev's periodic table led to the search for elements to fill the gaps in his chart. Morgan's years of research on chromosomes would have been virtually pointless were it not for the gene theory. The planet Mercury was discovered by pointing a telescope in a direction in which theory predicted such a mass must be.

Why, then, do accounts of scientific discovery emphasize serendipity so much? There are a number of possible reasons. The heuristic value of a theory may simply not be one of its strongest points. Furthermore, the accounts given are nearly always accounts of outstanding discoveries. But a discovery may be remembered as outstanding largely because it was unexpected and not predicted by any theory. Thus the scientific discoveries given detailed descriptions in the history books may be strongly biased in the direction of serendipitous ones. Such examples also tend to come from the newly emerging sciences. It would be difficult for theory to guide research in a discipline where no substantial theory has been developed. On the other hand, psychology is certainly not one of the well-developed sciences, and so it might well be expected to have much in common with less advanced sciences.

Key Ideas Box 14.2: The role of theory in guiding research

A *conventional description* of science is that implications are drawn from theories. The implications are called *hypotheses* and are put to experimental test. Verification of the hypothesis strengthens belief in the theory. If results do not fit the hypothesis, the classic view was that the theory is refuted. More recently, the view has been that theories are only abandoned when a better theory is available, even though the data do not always fit.

In fact, many if not most scientific discoveries occur *serendipitously.* This means that they occurred while the researcher was looking for something else. Thus the role of theory in aiding discovery has not been as great as the conventional description indicates. Further objection to theories as a guide to research comes from their tendency to *blind scientists to possibilities* outside the framework of the theory.

Nevertheless, theories are valuable. They *clarify* loose ideas. They sometimes give us good *ideas for research.* They help us *organize* and *remember* data. They also have a *predictive power* not provided by mere generalizations of functional relations. They often make predictions about quite novel situations, and they can predict the limiting conditions under which established functional relations will no longer hold.

Formal theories clarify loose verbal theories

We are stuck with theory whether we like it or not. Our only choice is whether to have a loose theory or a rigorous one. My students and I have often formalized theories of ours by doing computer simulations. I quickly learned that prior to the formal modeling, I simply did not know what I was talking about. I have never known anyone to formalize a theory without getting the same feeling. How can vague, inconsistent theories be better than clear, consistent ones? Clarification of ideas deepens understanding.

Informal theories inadvertently rely on the empathy of listeners. Equations and computers do not emphathize. Even a "theory" by so rugged an empiricist as Skinner, that presented in *Verbal Behavior* (1957), would require immense clarification before a computer could have a go at it. He knows what he means, I know what he means, you know what he means . . . or do we?

Study Questions

1. Explain the major methods of searching the scientific literature.
2. List the major questions to keep in mind as you read scientific literature.
3. Select an experiment (perhaps one from an earlier chapter) and show how you would follow up on it by verification, recombination, generalization, and specialization.
4. Explain the common sequence of development of a research topic from the stage of discovery to the stage of identifying underlying causes.
5. Distinguish between technique- and problem-oriented research.
6. What are the roles of other people as sources of ideas for research?
7. Give an example of an unverified belief that needs to be established scientifically. Try to make it your own instead of one from the text.
8. Explain the following:
   a. conventional description of science
   b. serendipity
9. What are the main advantages and disadvantages of theory in science?

15 COMMUNICATING THE RESULTS OF RESEARCH

The final stage in experimental psychology is communication. Science is inherently social. Remember the emphasis on agreement among observers early in this book. We can only be scientific by being objective, and to be objective is to have agreement between observers. So the very definition of a *fact* is, at bottom, a social one. Science begins with a social act, and it also ends in a social act—the act of communicating findings.

Many investigators balk at the stage of communication. They feel hesitant to communicate their findings to others in any systematic way. But this is to leave the task unfinished. The results of our work should be communicated to others. The skills necessary to do this are well within reach.

Even the work done while a student can be publishable. My impression is that students, especially undergraduates, commonly carry out projects that, with a little extra effort, could be published. Why not put in the extra effort? It is personally rewarding, and it makes the contribution of experimental subjects worthwhile.

Even if work is not to be published, it generally has to be described in writing as part of a course requirement. So you will likely have to learn to write in a style appropriate to psychological journals. The main

purpose of this chapter is to guide you in such writing, whether it stops with your teacher or goes on to a journal. I will also discuss some special problems involved in oral presentation of research.

Writing in APA style

There is a fairly strict set of requirements for writing in the style of the American Psychological Association. These requirements have been summarized in the *Publication Manual of the American Psychological Association,* 2nd ed. (1974). You should refer to that manual for details. Here, I will summarize highlights and also provide specific recommendations based on errors commonly made by students.

The general style used by APA has been developed under certain inevitable restraints. This accounts for some of its rigidity, and for the differences between "journal style" and the more breezy styles used in things like newspapers and magazines. Publication is costly, and therefore space is short for professional journals. The need for accuracy is, at the same time, very great. So a rather curt, concise style has been developed. And there is little toleration for deviations from that style.

TITLE

The title should summarize the main idea of the paper. One rule is to mention the independent variable, dependent variable, and a relationship between them. For example, "On extrinsic reward" is not good. It would be better to say "Effects of extrinsic reward on intrinsically motivated behaviors."

AUTHOR'S NAME AND INSTITUTIONAL AFFILIATION

Beneath the title list the name of the author or authors, without titles or degrees. Beneath that, list the institutional affiliation. Spell out the title of the institution. A common error is to use an abbreviation or nickname that is only locally understood. For example, "OSU" means "Ohio State University" if you are in the Midwest, but it means "Oklahoma State University" if you are in the Southwest. In foreign countries, it may mean nothing. You may feel that it is unlikely that your paper will reach Vladivostok, but the idea is to establish the right habits in case future papers of yours do get that far.

ABSTRACT

A brief summary (100–175 words) of the problem, method, results, and conclusions is next. The most important thing is to tell what you did

and what results you got. The abstract must be capable of being understood by itself. Since it is so short, you cannot ramble. Be succinct. The abstract is very important because it is all many people will ever read.

INTRODUCTION

State the problem, its relationship to past literature or theory, and show why you selected your research tactic to solve that problem. Your awareness of the pertinent literature should be obvious here, but without getting lengthy. You can refer to previous summary reviews. For classroom papers, if you had limited sources of literature, you should explain those limitations. If you have a novel problem, you can usually at least mention literature relevant to its components. For example, a study of the influence of alcohol on chess playing might have few, if any, predecessors. But there is a literature on the influence of alcohol on behavior. Incidentally, if you make claims in the introduction, or anywhere else in the paper, support them with references to literature sources. We cannot rely on common lore here. In science we are trying to improve on that.

METHOD

In the method section you should describe how the study was conducted. It should enable investigators to repeat your study. It customarily has the following subheadings: "Subjects," "Apparatus," and "Procedure." Under *Subjects* explain how many and what kind of subjects you employed. Also tell how they were sampled from the population of interest. Mention any variables that might influence your results, especially their generality. Answers to such things as "What strain of rat did you use?" "How old were the subjects?" "What were their socioeconomic backgrounds?" should be included if potentially relevant to your findings. Loss of subjects (by dropping out or otherwise) should also be explained here.

Under *Apparatus,* describe your apparatus systematically. If it is a standard apparatus describe it so that another person could obtain it. If you are going to publish in a real journal, or rehearse for such publication, keep in mind that the article may find its way to foreign readers. So do not be provincial in your identification of the apparatus. A novel apparatus will have to be described in some detail. Keep a constant perspective while describing it (imagine yourself standing in front of it, or on top of it). A drawing or photograph may be very helpful.

The *Procedure* subsection is a description of what you did to the subjects. It includes a description of the experimental design with its

methods of achieving control of important variables. Be sure to include descriptions of how each thing was measured, and enough detail for another person to replicate your study. On the other hand, be as economical as possible with words.

RESULTS

The next major section after "Method" is the section reporting results. A common error here is to begin with a report of outcomes of statistical analyses before presenting the data. First describe the main thrust of your results. Always describe the data (usually in figures and tables) before going on to demonstrate their reliability by such devices as statistical analyses. Since results are usually described in tables and figures, I will discuss their construction in some detail.

Constructing tables and figures

Tables and figures are useful both at the end of an experiment, when experimenters want to communicate their findings to others, and during an experiment, to make the findings intelligible to the experimenters themselves. There is, to be sure, a distinction between the quality of tables and figures constructed for private use and those constructed for public use; there can be some reduction in the clarity of labeling and in the attractiveness of the work when it is only for local use. However, avoid being too lax. Errors can easily creep in or important aspects of the data be missed if experimenters are careless about how they present the data to themselves or to other members of the research team.

Tables and figures are the very heart of a report. People often look at them before they look at accompanying text. Hence, the tables and figures should convey their message completely, as far as possible independently of the text. Labels should be clear; avoid mystifying the reader with unintelligible abbreviations. Both for yourself and for the reader, the data will be easier to understand if the labels and headings are immediately and directly intelligible. The use of arbitrary codes such as "Group A" and "Condition 1" requires the memorization of that code before the data can be interpreted. The interpretation of data is too important for us to permit any unnecessary obstacles. If abbreviations must be used, try developing them so that they readily suggest the thing they represent.

Tables and figures should be uncluttered. It is tempting to try putting all or most of the information from an experiment in one table or figure, because this has the potential of laying out the various results side by side for comparison. But the eye does not easily discern
and numbers when they are in a crowd.

When do we choose tables rather than figures? There are no ha

fast rules. In many cases data can be communicated in either way. Tables are to be preferred when exact numbers need to be put down. For this reason, and because of the greater expense of figures, the APA manual advocates giving preference to tables. However, in psychology, the error of measurement is commonly so great that it is relatively unimportant, even sometimes misleading, to put down highly exact numbers (say, out to several decimal places). In these cases figures will do nicely. On the other hand, larger quantities of information can sometimes be put in a table better than a figure.

Figures have the advantage of presenting trends and relationships in visible form. Deviation of a given datum is more likely to be seen in a figure than in a number. For personal use it is often best to construct both tables and figures for the same numbers. Afterwards, you can select the most successful method for presentation to others. Figures must, of course, be used to present pictorial information such as photos and drawings of apparatus or sections of brain tissue.

It is easy to overlook one of the major advantages to be gained from constructing tables and, especially, figures. The act of constructing them *forces the experimenter to get close to the data.* A datum unnoticed might just as well never have been gathered; much of the process of discovery takes place when the experimenter is tabulating and graphing the data.

Specific guidelines for the construction of tables have been laid down by Woodford (1968).

TABLE TITLE. Titles should be concise, with the key concepts emphasized by placing them at or near the beginning of the title. Detailed information goes in footnotes. The first table can bear the brunt of communicating to the reader; you can drop out some details in the titles of later tables. Avoid vague titles like "Results of experiment one," or long-winded ones like "The effects of hair length on the attractiveness of 40 male humans as measured in a natural setting in 312 female humans by the method of pupillary dilation." Instead try "Amount of pupillary dilation as a function of hair length". You should avoid unnecessary changes of pattern, such as changes from one table to the next in the order of the words in the title and headings.

COLUMN HEADINGS. Items should be grouped logically. Give control values first, in the far left column or in the top row. Data from other treatment conditions should be placed in their natural order, if one exists. For example, if you were reporting the results of an experiment comparing the effects on learning of three different levels of shock, then you could put your unshocked control results to the far left, place the results from the lowest shock level next to the controls, the intermediate level to the right of that, and the highest level to the far

right. Woodford (1968) recommends modifying this in the event that there is a particular pattern laid down in earlier tables that can be maintained in later ones. Readers might be better off with the familiar pattern instead of having to adjust to a change.

FOOTNOTES. More detailed information should go in footnotes. This particularly includes mention of statistical values. For example: "Long hair condition differed significantly from controls ($p < 0.02$)." Abbreviations are explained in footnotes. Here also, special conditions and qualifications are mentioned. For example "A 300-trial cut-off was employed," in an experiment using trials to a learning criterion as a dependent measure.

JOURNAL FORMAT. A final point about constructing tables is that they should be in the correct format for the place in which they are to be used. If this is a psychology journal, the format will probably be that of the American Psychological Association. The APA style is also commonly required by teachers in psychology classes.

In APA style, tables are typed, double-spaced, each on a separate page. The top heading is the word "Table" and its Arabic numeral (for example, "Table 1"). This is centered at the head of the table; then comes the title, centered, with principal words capitalized. Column and subheadings are centered with the initial letter of the first word capitalized. Footnotes are typed, paragraph indented, double-spaced, at the foot of the table.

FIGURES. I will restrict this discussion to the construction of datagrams. The word "figures" includes such things as photomicrographs of brain sections, which do not require discussion here.

As with tables, the figure should be a complete unit of communication. It consists of the figure itself with *labeled* axes: some sort of legend for indicating the meaning of the included bars, lines, and so on; a title; and a caption giving explanatory details. Titles and captions are typewritten, double-spaced, on a separate sheet, not placed on the figure itself. (This rule may be modified for classroom purposes.)

If the figure is to be used for publication, you should consider how it will look in the format of the journal. It makes a difference whether the journal is physically large or not, for example. People often make the lettering on their figures so small that it becomes unreadable once the figure has been reduced to journal size. Color printing is generally not available, so differentiations cannot be made by using different colors.

Many different dimensions can be used to indicate differences in a datagram. We usually indicate differences in the variables of interest by differences in the height or length of bars or points on a graph. By

paging through this book you can see many different examples of reasonably well made figures.

Datagrams should not be cluttered. Usually no more than four curves should appear on a given graph. Labeling should be very clear. Avoid using abbreviations if possible, and if they are necessary, try to make them easily related to the full words. It is often helpful to use meaningful coding symbols instead of dots on a line-graph. For example, a datagram showing the effects of some treatment on male and female subjects might use ♂ and ♀ instead of dots for the two groups. Letters of the alphabet can also be used in this way. For example, I have used little H's and V's instead of dots to indicate treatments where an animal was learning to choose a horizontal or vertical cue. Another aid to understanding can be provided if labels identifying the various lines and bars are placed near them, with arrows specifying which label goes with which bar or graph.

Accuracy of data in tables and datagrams

Before finishing this discussion, it is important to point out that errors often crop up in transcribing data from raw form to tables and datagrams. You should always check the data against your end product very carefully. Recalculating totals of columns of numbers from the typescript is often worthwhile. Also, check for discrepancies when tabulated or graphed numbers are cited in the text.

Reliable and unreliable results

Experimenters sometimes report mean differences between conditions that are unreliable by their stipulated test (such as the 0.05 level), but then go on to treat the differences as though they were confirmed. This is not legitimate. Keep in mind that the test for reliability of the data tells us whether or not we have reason to suppose that there is anything but error of measurement underlying them. If we fail to reject the null hypothesis, we have no reason to suppose that there is anything more underlying observed differences in central tendency than underlies a good poker hand. It is unscientific to infer trends from results that "almost meet the 0.05 level."

DISCUSSION

Here you discuss the implications of your data. A common error is to forget about the statement of the original problem when it is time for the discussion. You should start by saying how your results relate to the original problem posed. Another common error is to turn this section into a kind of confession of errors. It is not necessary to obsessively point out every conceivable flaw in your methods. A third common error is to ignore your findings. This may take the form of

adhering to a previous theoretical predilection despite your data's having indicated the contrary. Another form it takes is to disparage your results because they do not seem to agree with other, published findings. It is not out of the question that the other investigators had flaws in their experiments.

You should accept your findings, relate them to your introductory statements, and draw out their theoretical and practical implications, as well as their implications for future research.

REFERENCES

Throughout the paper, statements should be supported by references. The usual style in the text is to refer to such things as ". . . the pioneering study of Jones (1975) . . ." or "There have been many studies on chess playing in chimpanzees (Jones, 1975; Smith, 1976)." At the end of the paper a reference list is provided. It includes only papers cited in the text. There is a standard APA format for citations. Begin with a list of all authors of the work, last name first, followed by initials. Next put the title of the article, chapter, or book. For journals, next put the name of the journal in full, the date of publication, the volume number, and the first and last page numbers. For books, next put the city of publication, the publisher's name, and the date of publication. Articles and books referenced should be listed in alphabetical order of the first author's last name.

APPENDIX

Appendices are sometimes permitted by journal editors. These contain material that would be distracting in the main body of the text, but would be useful in understanding, evaluating, or replicating your study. Appendices are quite commonly used in papers presented as requirements for class and in theses. In these cases they often include raw data presented in an orderly and intelligible way. You should find out whether your teacher wants you to include such appendices.

Oral presentations

Second only to published articles in communicating the results of research is the oral presentation. People often follow the same format as the journals require. This is usually a source of suffering for the listeners, who generally find such presentations boring and incomprehensible.

There are a number of major differences between the oral and the written mode. When a written article is available, the reader can reread points not remembered or perhaps not even comprehended. No such

opportunity exists with the oral mode. Generally oral presentations are done at conventions where time is severely limited (often 10–20 min), and little time is available for questioning. So it is best to depart from journal format and present things in a more breezy style with a great deal of redundancy.

One of the best techniques I have ever seen was one used by Ronald Myers (the "brain-splitter" of Chapter 11). He presented not only the customary slides summarizing his data, but also a series of slides that summarized the main points made in his presentation up to that point. This way, if something was missed, the main ideas were repeated at regular intervals and could be recaptured. Perhaps you will not go that far (then again, why not?) but at least you can repeat your main points in one form or another at various stages of your presentation. Incidentally, be sure the details on your slides are large enough to be seen at the back of a lecture hall. Remarkably often slides presented at conventions are well below the normal threshold of acuity!

Another major difference between oral and written presentation is the tendency for fear to impair oral performance. Close to 90 percent of us fear speaking in front of groups. The symptoms of that fear can be virtually incapacitating. The best way to deal with it is to realize that nearly everyone has to go through the same thing, and therefore virtually everyone will understand if you display the symptoms. Do not feel that you lose face by showing nervousness in front of a group. In fact, it is far better to simply accept that nervousness and try to present the *facts* in spite of it.

Notice my emphasis on presenting the facts. You are there for that purpose, and your audience is there to hear those facts, not to evaluate you. If you get the facts across, you have done better than many people who use journal format and illegible slides, and leave their listeners in the dark. You can arrange to make sure the facts get across by putting them on eminently clear slides. In fact, if you are nervous it is sometimes best to begin with a slide. That will get you started, and it is relatively easy to describe a familiar slide (being in the dark may help, too!).

You may spend a lot of time before the presentation anticipating all kinds of incisive criticisms. This is good up to a point. But only up to a point. Make yourself aware of potential criticisms, but do not get repetitious about it. Remember, you know your experiment better than anyone else. Chances are that the audience will be preoccupied with understanding what you did and what you got for results. They will hardly have time to think through the study better than you did, if you are at all careful. Unfortunately (or fortunately, depending on your point of view), there is usually little time for discussion and questioning anyway.

The only way you get over fear of oral presentations is through

Key Ideas Box 15.1: Communication

Communication of data is an essential aspect of the scientific enterprise. It is most often done in *journal format.* Journal articles are composed of:

1. *Title:* usually including statement of functional relationship between independent and dependent variables.
2. *Author's name and affiliation.*
3. *Abstract:* a 100–175 word summary of what you did and what you got.
4. *Introduction:* it is not labeled, but includes a statement of the problem, the relevant literature, and the relationship of your study to that problem.
5. *Method:* includes (a) *Subjects* (how sampled and any other relevant information); (b) *Apparatus* (described systematically, so it could be reproduced or obtained); (c) *Procedure,* which tells what you did to the subjects, including experimental design and methods of measurement.
6. *Results,* in which you describe your data, usually in tables or figures, and tell how you assessed their reliability.
7. *Discussion,* in which you tell how the results relate to the originally stated problem and give other implications for theory and future research.
8. *References,* in which you list, in alphabetical order by last name of first author, all articles referred to in the paper (and no other articles).
9. *Appendix:* it is only occasionally used, but contains detailed material that would be distracting in the main body of the article.

Oral presentations should not be in journal format, though they should provide most of the relevant details. They differ from written presentations in that they require repetitiousness and overcoming of fear. The text describes methods of achieving redundancy and of reducing fear.

practicing them. Avoidance behaviors can last a very long time, even when the aversive stimulus is not there. So take every opportunity to sharpen your ability to present data in this way. Sometimes teachers will let you rehearse in class. Be the first volunteer.

If your problems in front of an audience overwhelm you, try special relaxation procedures such as autogenic training (Lindemann, 1973), progressive relaxation (Jacobson, 1957), or self-hypnosis (Anderson & Savary, 1972). If necessary, find someone to give you training in

desensitization. Above all, don't give up on extinguishing the responses.

Study Questions

1. Outline and briefly describe the components of an article written in APA style.
2. Why not infer trends from almost significant results?
3. Outline the major characteristics of a good table.
4. Outline the major characteristics of a good figure.
5. What are the main differences between written and oral presentations?

APPENDICES

APPENDIX A Table of Random Numbers

	COLUMN NUMBER							
Row	00000 01234	00000 56789	11111 01234	11111 56789	22222 01234	22222 56789	33333 01234	33333 56789
				1st Thousand				
00	23157	54859	01837	25993	76249	70886	95230	36744
01	05545	55043	10537	43508	90611	83744	10962	21343
02	14871	60350	32404	36223	50051	00322	11543	80834
03	38976	74951	94051	75853	78805	90194	32428	71695
04	97312	61718	99755	30870	94251	25841	54882	10513
05	11742	69381	44339	30872	32797	33118	22647	06850
06	43361	28859	11016	45623	93009	00499	43640	74036
07	93806	20478	38268	04491	55751	18932	58475	52571
08	49540	13181	08429	84187	69538	29661	77738	09527
09	36768	72633	37948	21569	41959	68670	45274	83880
10	07092	52392	24627	12067	06558	45344	67338	45320
11	43310	01081	44863	80307	52555	16148	89742	94647
12	61570	06360	06173	63775	63148	95123	35017	46993
13	31352	83799	10779	18941	31579	76448	62584	86919
14	57048	86526	27795	93692	90529	56546	35065	32254
15	09243	44200	68721	07137	30729	75756	09298	27650
16	97957	35018	40894	88329	52230	82521	22532	61587
17	93732	59570	43781	98885	56671	66826	95996	44569
18	72621	11225	00922	68264	35666	59434	71687	58167
19	61020	74418	45371	20794	95917	37866	99536	19378
20	97839	85474	33055	91718	45473	54144	22034	23000
21	89160	97192	22232	90637	35055	45489	88438	16361
22	25966	88220	62871	79265	02823	52862	84919	54883
23	81443	31719	05049	54806	74690	07567	65017	16543
24	11322	54931	42362	34386	08624	97687	46245	23245

Reprinted with permission from M. G. Kendall and B. B. Smith, Randomness and random sampling numbers. *Journal of the Royal Statistical Society*, 1938, *101*.

APPENDIX B Table of Binomial Probabilities

N \ *X*	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
5	031	188	500	812	969	*										
6	016	109	344	656	891	984	*									
7	008	062	227	500	773	938	992	*								
8	004	035	145	363	637	855	965	996	*							
9	002	020	090	254	500	746	910	980	998	*						
10	001	011	055	172	377	623	828	945	989	999	*					
11		006	033	113	274	500	726	887	967	994	*	*				
12		003	019	073	194	387	613	806	927	981	997	*	*			
13		002	011	046	133	291	500	709	867	954	989	998	*	*		
14		001	006	029	090	212	395	605	788	910	971	994	999	*	*	
15			004	018	059	151	304	500	696	849	941	982	996	*	*	*
16			002	011	038	105	227	402	598	773	895	962	989	998	*	*
17			001	006	025	072	166	315	500	685	834	928	975	994	999	*
18			001	004	015	048	119	240	407	593	760	881	952	985	996	999
19				002	010	032	084	180	324	500	676	820	916	968	990	998
20				001	006	021	058	132	252	412	588	748	868	942	979	994
21				001	004	013	039	095	192	332	500	668	808	905	961	987
22					002	008	026	067	143	262	416	584	738	857	933	974
23					001	005	017	047	105	202	339	500	661	798	895	953
24					001	003	011	032	076	154	271	419	581	729	846	924
25						002	007	022	054	115	212	345	500	655	788	885

APPENDIX C Table of Critical Values for Mann-Whitney U

$n_2 = 3$

U \ n_1	1	2	3
0	.250	.100	.050
1	.500	.200	.100
2	.750	.400	.200
3		.600	.350
4			.500
5			.650

$n_2 = 4$

U \ n_1	1	2	3	4
0	.200	.067	.028	.014
1	.400	.133	.057	.029
2	.600	.267	.114	.057
3		.400	.200	.100
4		.600	.314	.171
5			.429	.243
6			.571	.343
7				.443
8				.557

$n_2 = 5$

U \ n_1	1	2	3	4	5
0	.167	.047	.018	.008	.004
1	.333	.095	.036	.016	.008
2	.500	.190	.071	.032	.016
3	.667	.286	.125	.056	.028
4		.429	.196	.095	.048
5		.571	.286	.143	.075
6			.393	.206	.111
7			.500	.278	.155
8			.607	.365	.210
9				.452	.274
10				.548	.345
11					.421
12					.500
13					.579

$n_2 = 6$

U \ n_1	1	2	3	4	5	6
0	.143	.036	.012	.005	.002	.001
1	.286	.071	.024	.010	.004	.002
2	.428	.143	.048	.019	.009	.004
3	.571	.214	.083	.033	.015	.008
4		.321	.131	.057	.026	.013
5		.429	.190	.086	.041	.021
6		.571	.274	.129	.063	.032
7			.357	.176	.089	.047
8			.452	.238	.123	.066
9			.548	.305	.165	.090
10				.381	.214	.120
11				.457	.268	.155
12				.545	.331	.197
13					.396	.242
14					.465	.294
15					.535	.350
16						.409
17						.469
18						.531

Reprinted with permission from H. B. Mann and D. R. Whitney, On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.*, 1947, *18*, 52–54.

Table of Critical Values for Mann-Whitney U (continued)

$n_2 = 7$

U \ n_1	1	2	3	4	5	6	7
0	.125	.028	.008	.003	.001	.001	.000
1	.250	.056	.017	.006	.003	.001	.001
2	.375	.111	.033	.012	.005	.002	.001
3	.500	.167	.058	.021	.009	.004	.002
4	.625	.250	.092	.036	.015	.007	.003
5		.333	.133	.055	.024	.011	.006
6		.444	.192	.082	.037	.017	.009
7		.556	.258	.115	.053	.026	.013
8			.333	.158	.074	.037	.019
9			.417	.206	.101	.051	.027
10			.500	.264	.134	.069	.036
11			.583	.324	.172	.090	.049
12				.394	.216	.117	.064
13				.464	.265	.147	.082
14				.538	.319	.183	.104
15					.378	.223	.130
16					.438	.267	.159
17					.500	.314	.191
18					.562	.365	.228
19						.418	.267
20						.473	.310
21						.527	.355
22							.402
23							.451
24							.500
25							.549

Table of Critical Values for Mann-Whitney U (continued)

$n_2 = 8$

U \ n_1	1	2	3	4	5	6	7	8	t	Normal
0	.111	.022	.006	.002	.001	.000	.000	.000	3.308	.001
1	.222	.044	.012	.004	.002	.001	.000	.000	3.203	.001
2	.333	.089	.024	.008	.003	.001	.001	.000	3.098	.001
3	.444	.133	.042	.014	.005	.002	.001	.001	2.993	.001
4	.556	.200	.067	.024	.009	.004	.002	.001	2.888	.002
5		.267	.097	.036	.015	.006	.003	.001	2.783	.003
6		.356	.139	.055	.023	.010	.005	.002	2.678	.004
7		.444	.188	.077	.033	.015	.007	.003	2.573	.005
8		.556	.248	.107	.047	.021	.010	.005	2.468	.007
9			.315	.141	.064	.030	.014	.007	2.363	.009
10			.387	.184	.085	.041	.020	.010	2.258	.012
11			.461	.230	.111	.054	.027	.014	2.153	.016
12			.539	.285	.142	.071	.036	.019	2.048	.020
13				.341	.177	.091	.047	.025	1.943	.026
14				.404	.217	.114	.060	.032	1.838	.033
15				.467	.262	.141	.076	.041	1.733	.041
16				.533	.311	.172	.095	.052	1.628	.052
17					.362	.207	.116	.065	1.523	.064
18					.416	.245	.140	.080	1.418	.078
19					.472	.286	.168	.097	1.313	.094
20					.528	.331	.198	.117	1.208	.113
21						.377	.232	.139	1.102	.135
22						.426	.268	.164	.998	.159
23						.475	.306	.191	.893	.185
24						.525	.347	.221	.788	.215
25							.389	.253	.683	.247
26							.433	.287	.578	.282
27							.478	.323	.473	.318
28							.522	.360	.368	.356
29								.399	.263	.396
30								.439	.158	.437
31								.480	.052	.481
32								.520		

Table of Critical Values for Mann-Whitney U (continued)

Critical Values of U for a One-tailed Test at $\alpha = .001$ or for a Two-tailed Test at $\alpha = .002$

n_1 \ n_2	9	10	11	12	13	14	15	16	17	18	19	20
1												
2												
3									0	0	0	0
4		0	0	0	1	1	1	2	2	3	3	3
5	1	1	2	2	3	3	4	5	5	6	7	7
6	2	3	4	4	5	6	7	8	9	10	11	12
7	3	5	6	7	8	9	10	11	13	14	15	16
8	5	6	8	9	11	12	14	15	17	18	20	21
9	7	8	10	12	14	15	17	19	21	23	25	26
10	8	10	12	14	17	19	21	23	25	27	29	32
11	10	12	15	17	20	22	24	27	29	32	34	37
12	12	14	17	20	23	25	28	31	34	37	40	42
13	14	17	20	23	26	29	32	35	38	42	45	48
14	15	19	22	25	29	32	36	39	43	46	50	54
15	17	21	24	28	32	36	40	43	47	51	55	59
16	19	23	27	31	35	39	43	48	52	56	60	65
17	21	25	29	34	38	43	47	52	57	61	66	70
18	23	27	32	37	42	46	51	56	61	66	71	76
19	25	29	34	40	45	50	55	60	66	71	77	82
20	26	32	37	42	48	54	59	65	70	76	82	88

Critical Values of U for a One-tailed Test at $\alpha = .01$ or for a Two-tailed Test at $\alpha = .02$

n_1 \ n_2	9	10	11	12	13	14	15	16	17	18	19	20
1												
2					0	0	0	0	0	0	1	1
3	1	1	1	2	2	2	3	3	4	4	4	5
4	3	3	4	5	5	6	7	7	8	9	9	10
5	5	6	7	8	9	10	11	12	13	14	15	16
6	7	8	9	11	12	13	15	16	18	19	20	22
7	9	11	12	14	16	17	19	21	23	24	26	28
8	11	13	15	17	20	22	24	26	28	30	32	34
9	14	16	18	21	23	26	28	31	33	36	38	40
10	16	19	22	24	27	30	33	36	38	41	44	47
11	18	22	25	28	31	34	37	41	44	47	50	53
12	21	24	28	31	35	38	42	46	49	53	56	60
13	23	27	31	35	39	43	47	51	55	59	63	67
14	26	30	34	38	43	47	51	56	60	65	69	73
15	28	33	37	42	47	51	56	61	66	70	75	80
16	31	36	41	46	51	56	61	66	71	76	82	87
17	33	38	44	49	55	60	66	71	77	82	88	93
18	36	41	47	53	59	65	70	76	82	88	94	100
19	38	44	50	56	63	69	75	82	88	94	101	107
20	40	47	53	60	67	73	80	87	93	100	107	114

Adapted and abridged from Tables 1, 3, 5, and 7 of D. Auble, Extended tables for the Mann-Whitney statistic. *Bulletin of the Institute of Educational Research at Indiana University*, 1953, *1*, No. 2.

Table of Critical Values for Mann-Whitney U (continued)

Critical Values of U for a One-tailed Test at α = .025 or for a Two-tailed Test at α = .05

n_1 \ n_2	9	10	11	12	13	14	15	16	17	18	19	20
1												
2	0	0	0	1	1	1	1	1	2	2	2	2
3	2	3	3	4	4	5	5	6	6	7	7	8
4	4	5	6	7	8	9	10	11	11	12	13	13
5	7	8	9	11	12	13	14	15	17	18	19	20
6	10	11	13	14	16	17	19	21	22	24	25	27
7	12	14	16	18	20	22	24	26	28	30	32	34
8	15	17	19	22	24	26	29	31	34	36	38	41
9	17	20	23	26	28	31	34	37	39	42	45	48
10	20	23	26	29	33	36	39	42	45	48	52	55
11	23	26	30	33	37	40	44	47	51	55	58	62
12	26	29	33	37	41	45	49	53	57	61	65	69
13	28	33	37	41	45	50	54	59	63	67	72	76
14	31	36	40	45	50	55	59	64	67	74	78	83
15	34	39	44	49	54	59	64	70	75	80	85	90
16	37	42	47	53	59	64	70	75	81	86	92	98
17	39	45	51	57	63	67	75	81	87	93	99	105
18	42	48	55	61	67	74	80	86	93	99	106	112
19	45	52	58	65	72	78	85	92	99	106	113	119
20	48	55	62	69	76	83	90	98	105	112	119	127

Critical Values of U for a One-tailed Test at α = .05 or for a Two-tailed Test at α = .10

n_2 \ n_1	9	10	11	12	13	14	15	16	17	18	19	20
1											0	0
2	1	1	1	2	2	2	3	3	3	4	4	4
3	3	4	5	5	6	7	7	8	9	9	10	11
4	6	7	8	9	10	11	12	14	15	16	17	18
5	9	11	12	13	15	16	18	19	20	22	23	25
6	12	14	16	17	19	21	23	25	26	28	30	32
7	15	17	19	21	24	26	28	30	33	35	37	39
8	18	20	23	26	28	31	33	36	39	41	44	47
9	21	24	27	30	33	36	39	42	45	48	51	54
10	24	27	31	34	37	41	44	48	51	55	58	62
11	27	31	34	38	42	46	50	54	57	61	65	69
12	30	34	38	42	47	51	55	60	64	68	72	77
13	33	37	42	47	51	56	61	65	70	75	80	84
14	36	41	46	51	56	61	66	71	77	82	87	92
15	39	44	50	55	61	66	72	77	83	88	94	100
16	42	48	54	60	65	71	77	83	89	95	101	107
17	45	51	57	64	70	77	83	89	96	102	109	115
18	48	55	61	68	75	82	88	95	102	109	116	123
19	51	58	65	72	80	87	94	101	109	116	123	130
20	54	62	69	77	84	92	100	107	115	123	130	138

APPENDIX D Table of Percentile Points for t

df	*one-tailed* $p = 0.4$	*0.25*	*0.1*	*0.05*	*0.025*	*0.01*	*0.005*	*0.001*
	two-tailed $p = 0.8$	*0.5*	*0.2*	*0.1*	*0.05*	*0.02*	*0.01*	*0.002*
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	318.31
2	.289	0.816	1.886	2.920	4.303	6.965	9.925	22.326
3	.277	.765	1.638	2.353	3.182	4.541	5.841	10.213
4	.271	.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	.265	.718	1.440	1.943	2.447	3.143	3.707	5.208
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.785
8	.262	.706	1.397	1.860	2.306	2.896	3.355	4.501
9	.261	.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	4.144
11	.260	.697	1.363	1.796	2.201	2.718	3.106	4.025
12	.259	.695	1.356	1.782	2.179	2.681	3.055	3.930
13	.259	.694	1.350	1.771	2.160	2.650	3.012	3.852
14	.258	.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16	.258	.690	1.337	1.746	2.120	2.583	2.921	3.686
17	.257	.689	1.333	1.740	2.110	2.567	2.898	3.646
18	.257	.688	1.330	1.734	2.101	2.552	2.878	3.610
19	.257	.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.527
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.505
23	.256	.685	1.319	1.714	2.069	2.500	2.807	3.485
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.450
26	.256	.684	1.315	1.706	2.056	2.479	2.779	3.435
27	.256	.684	1.314	1.703	2.052	2.473	2.771	3.421
28	.256	.683	1.313	1.701	2.048	2.467	2.763	3.408
29	.256	.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.385
40	.255	.681	1.303	1.684	2.021	2.423	2.704	3.307
60	.254	.679	1.296	1.671	2.000	2.390	2.660	3.232
120	.254	.677	1.289	1.658	1.980	2.358	2.617	3.160
∞	.253	.674	1.282	1.645	1.960	2.326	2.576	3.909

APPENDIX E Table of Percentile Points for *F*

$\alpha = 0.05$

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Abridged from Table 18 of the *Biometrika Tables for Statisticians*, Vol. I (ed. 3), 1966, edited by E. S. Pearson and H. O. Hartley.

$\alpha = 0.025$

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	976.7	984.9	993.1	997.2	1001	1006	1010	1014	1018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00

$\alpha = 0.01$

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	4999.5	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	8.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

References

Ad Hoc Committee on Ethical Standards in Psychological Research. *Ethical principles in the conduct of research with human participation.* Washington, D.C.: American Psychological Association, 1973, pp. 1–2.

Alrutz, S. Studien auf dem gebiete der Temperatursinne, *Skandinavisches Archiv Für Physiologie,* 1897, **7**, 321–340.

American Psychological Association. Publication manual of the American Psychological Association, 2nd ed. Washington, D.C., 1974.

Anderson, M. S., & Savary, I. M. *Passages: A guide for pilgrims of the mind.* New York, Harper & Row, 1972.

Andrade, E. N. DaC. *Sir Isaac Newton.* Garden City, N.Y.: Doubleday, 1958.

Anrep, G. V. Pitch discrimination in the dog, *Journal of Physiology,* 1920, **53**, 367–385.

Ardrey, R. *African genesis.* New York: Atheneum, 1961.

Aserinsky, E., & Kleitman, N. Regularly occurring periods of eye motility and concomitant phenomena during sleep, *Science,* 1953, **118**, 273–274.

Aserinsky, E., & Kleitman, N. Two types of ocular motility occurring in sleep, *Journal of Applied Physiology,* 1955, **8**, 1–10.

Atkinson, J. W., & McClelland, D. C. The projective expression of needs. II. The effect of different intensities of the hunger drive on thematic apperception, *Journal of Experimental Psychology,* 1948, **38**, 643–658.

Averbach, E., & Coriell, A. Short-term memory in vision, *Bell System Technical Journal,* 1961, **40**, 309–328.

Banks, W. Signal detection theory and human memory, *Psychological Bulletin,* 1970, **74,** 81–99.

Barber, T. X. *Hypnosis, a scientific approach.* NewYork: Van Nostrand, 1969.

Barber, T. X., & Hahn, K. W., Jr. Physiological and subjective responses to pain-producing stimulation under hypnotically-suggested and waking-imagined "analgesia," *Journal of Abnormal and Social Psychology,* 1962, **65,** 411–418.

Barber, T. X., & Silver, M. J. Fact, fiction and the experimenter bias effect, *Psychological Bulletin,* 1968, **70,** (6, Pt. 2), 1–29.

Barber, T. X., Spanos, N., & Chaves, J. *"Hypnosis," imagination, and human potentialities.* Elmsford, N.Y.: Pergamon, 1974, Chapter 10.

Bartlett, F. H. *Remembering: An experimental and social study.* London: Cambridge University Press, 1932.

Beech, H. R. *Changing man's behavior.* Baltimore: Penguin, 1969.

Beecher, H. K. *Measurement of subjective responses: Quantitative effects of drugs.* New York: Oxford University Press, 1959.

Beecher, H. K. Increased stress and effectiveness of placebos and "active" drugs, *Science,* 1960, **132,** 91–92.

Bekesy, G. von. Similarities between hearing and skin sensations, *Psychological Review,* 1959, **66,** 1–22.

Bellack, A. S. A comparison of self-reinforcement and self-monitoring in a weight reduction program, *Behavior Therapy,* in press.

Bellack, A. S., Rozensky, R., & Schwartz, J. A comparison of two forms of self-monitoring in a behavioral weight reduction program, *Behavior Therapy,* 1974, **5,** 523–530.

Bellack, A. S., Schwartz, J., & Rozensky, R. H. The contribution of external control to self-control in a weight reduction program, *Behavior Therapy and Experimental Psychiatry,* 1974, **5,** 245–249.

Berger, R., & Meier, G. Deprivation of patterned vision and the eye movements of sleep, *APSS,* 1965, Washington, D.C. Cited in Luce, 1966.

Bermant, G., & Starr, M. On telling people what they are likely to do, *Proceedings of the 80th Annual Convention of the American Psychological Association,* 1972, Vol. 7, p. 131.

Berne, E. *Games people play: The psychology of human relations.* New York: Grove Press, 1964.

Bertalanffy, L. von. *General systems theory.* NewYork: Braziller, 1968.

Besley, S., & Sheridan, C. L. A sensitive method for the detection of experimental brain lesions, *Perceptual and Motor Skills,* 1973, **36,** 584–586.

Best, J. B. Protopsychology, *Scientific American,* 1963, **208** (2), 55–62.

Bettinger, L. A., Danis, J. L., Meikle, M. B., Birch, H., Knopp, R., Smith, H. E., & Thompson, R. E. "Novelty" cells in association cortex of cat, *Psychonomic Science,* 1967, **9,** 421–422.

Beveridge, W. I. B. *The art of scientific investigation.* New York: Random House (Vintage), 1950.

Bexton, W. H., Heron, W., & Scott, T. H. Effects of decreased variation in the sensory environment, *Canadian Journal of Psychology,* 1954, **8,** 70–76.

Bianki, V. L. The role of the corpus callosum in paired activity of the optical and cutaneous analyzers in the rabbit, *Pavlovian Journal of Higher Nervous Activity,* 1959, **9,** 110–120.

Bickman, L. Social roles and uniforms: Clothes make the person, *Psychology Today,* April 1974, **7** (11), 49–51.

Bilbrey, J., & Winokur, S. Controls for and constraints on autoshaping, *Journal of the Experimental Analysis of Behavior,* 1973, **20**, 323–332.

Binet, A., & Simon, T. *The development of intelligence in children.* (Kite, E. S., Trans.) Baltimore: Williams & Wilkins, 1916.

Birch, H. G. The relation of previous experience to insightful problem solving, *Journal of Comparative and Physiological Psychology,* 1945, **38**, 367–383.

Birren, F. The emotional significance of color preference, *American Journal of Occupational Therapy,* 1952, **6**, 61–88.

Bishop, H. E. Innateness and learning in the visual perception of direction. Unpublished doctoral dissertation, University of Chicago, 1959.

Bjorn, H. Elektrisk Retning av Tanderoch dess Tillampning inom Tandlakarkonste, *Svensk Tandlakare-Tidskrift,* 1946, **39**, 625–635.

Black, P., & Myers, R. Visual function of the forebrain commissures in the chimpanzee, *Science,* 1964, **146**, 799–800.

Blough, D. S. Dark adaptation in the pigeon, *Journal of Comparative and Physiological Psychology,* 1956, **49**, 425–430.

Blum, R. A., & Blum, J. S. Factual issues in the continuity controversy, *Psychological Review,* 1949, **56**, 33–50.

Boring, E. G. *Sensation and perception in the history of experimental psychology.* New York: Appleton, 1942.

Boring, E. G. *A history of experimental psychology,* 2nd ed. New York: Appleton, 1950.

Bornstein, P., & Siprelle, C. Induced anxiety in the treatment of obesity: A preliminary case report, *Behavior Therapy,* 1973, **4**, 141–143.

Brady, J. V. Ulcers in "executive" monkeys, *Scientific American,* 1958, **199** (4), 95–100.

Braun, J. J., Meyer, P. M., & Meyer, D. R. Sparing of a brightness habit in rats following visual decortication, *Journal of Comparative and Physiological Psychology,* 1966, **61**, 79–82.

Breland, K., & Breland, M. The misbehavior of organisms, *American Psychologist,* 1961, **16**, 681–684.

Broadbent, D. A mechanical model for human attention and immediate memory, *Psychological Review,* 1957, **64**, 205–215.

Brown, J. Factors affecting self-punitive locomotor behavior. In Campbell, B., & Church, R. (Eds.), *Punishment and aversive behavior.* New York: Appleton, 1969.

Brown, J., Martin, R., & Morrow, M. Self-punitive behavior in the rat: Facilitative effects of punishment on resistance to extinction, *Journal of Comparative and Physiological Psychology,* 1964, **57**, 127–133.

Brown, P., & Jenkins, H. Auto-shaping of the pigeon's key-peck, *Journal of the Experimental Analysis of Behavior,* 1968, **11**, 1–8.

Bruning, J., & Kintz, B. *Computational handbook of statistics.* Glenview, Ill.: Scott, Foresman, 1968.

Bures, J., & Buresova, O. The reunified split brain. In Whalen, R. (Ed.), *Interdisciplinary conference on the neural control of behavior.* New York: Academic Press, 1970, pp. 211–238.

Donders, F. On the speed of mental processes (1869). In Koster, W. G. (Ed.), Attention and performance II, *Acta Psychologica,* 1969, **30**, 412–431.

Doob, L. Eidetic images among the Ibo, *Ethnology,* 1964, **3**, 357–363.

Downer, J. L. de C. Interhemispheric integration in the visual system. In V. B. Mountcastle (Ed.), *Interhemispheric relations and cerebral dominance.* Baltimore: Johns Hopkins Press, 1962, 87–100.

Drucker, E., Cannon, L., & Ware, J. The effects of sleep deprivation on performance over a 48-hour period. Fort Knox, Ky.: Human Resources Research Office, Division No. 2 (Armor), Tech. Rep. ITR-70-16. 1970.

Ebbinghaus, H. *Memory, a contribution to experimental psychology.* (1885) (Ruger, A. A., & Bussenius, C. E., Trans.) New York: Dover, 1964.

Edwards, D., & Edwards, J. Fetal movement: Development and time course, *Science,* 1970, **169**, 95–97.

Ehrenfreund, D. An experimental test of the continuity theory of discrimination learning with pattern vision, *Journal of Comparative and Physiological Psychology,* 1948, **41**, 408–422.

Ellison, D. Hallucinations produced by sensory conditioning, *Journal of Experimental Psychology,* 1941, **28**, 1–20.

Engen, T. Psychophysics II. Scaling methods. In Kling, J., & Riggs, L. (Eds.), *Woodworth and Schoosberg's Experimental Psychology,* 3rd. ed. New York: Holt, Rinehart and Winston, 1971, pp. 49–86.

Eysenck, Hans J. The effects of psychotherapy, an evaluation, *Journal of Consulting and Clinical Psychology,* 1952, **16**, 319–323.

Fantz, R. L. Ontogeny of perception. In Schrier, A. M., Harlow, H. F., & Stollnitz, F. (Eds.), *Behavior of nonhuman primates,* Vol. II. New York: Academic Press, 1965.

Fantz, R. L. Visual perception and experience in early infancy: A look at the hidden side of behavior development. In Stevenson, H. W., Hess, E. H., & Rheingold, H. L. (Eds.), *Early behavior.* New York: Wiley, 1967.

Fechner, G. T. *Elemente der Psychophysik.* (1860) Howes, D. H., & Boring, Edwin G. (Eds.). New York: Holt, Rinehart and Winston, 1966.

Feldt, L. S. A comparison of the precision of three experimental designs employing a concomitant variable, *Psychometrika,* 1958, **23**, 335–353.

Ferster, C., & Skinner, B. F. *Schedules of Reinforcement.* New York: Appleton, 1957.

Feynman, R. P., Leighton, R. B., & Sands, M. *The Feynman lectures on physics,* Vol. 1. Reading, Mass.: Addison-Wesley, 1963. Sect. 7.

Fields, P. E. Contributions to visual figure discrimination in the white rat, *Journal of Comparative Psychology,* 1931, **11**, 327–348.

Fishbein, W., Schaumberg, H., & Weitzmann, E. D. Rapid eye movements during sleep in dark reared kittens, *APSS,* 1965, Washington, D.C. Cited in Luce, 1966.

Fisher, S. *The female orgasm.* New York: Basic Books, 1973.

Folgmann, E. An experimental study of composer preferences of four outstanding orchestras, *Journal of Experimental Psychology,* 1933, **16**, 709–724.

Foulkes, D. Dream reports from different stages of sleep, *Journal of Abnormal and Social Psychology,* 1962, **65**, 14–25.

Foulkes, D. *The psychology of sleep.* New York: Scribner, 1966.

Frey, M. von Beiträge zur Sinnesphysiologie der Haut, *Ber. Sächs. Ges. Wiss,* 1895, **47**, 166–184. Cited in Boring, E. G., *Sensation and perception in the history of experimental psychology.* New York: Appleton, 1942. (P. 514 ref.)

Galanter, E. Contemporary psychophysics. In *New directions in psychology I.* New York: Holt, Rinehart, and Winston, 1962.

Galton, F. *Inquiries into human faculty and its development.* London: Macmillan, 1883.

Garcia, J., & Ervin, F. R. Gustatory-visceral and telereceptor-cutaneous conditioning—adaptation in internal milieus, *Communications in Behavioral Biology,* Part A, 1968, **1**, 389–415.

Garrett, H. E. *Great experiments in psychology.* New York: Appleton, 1951.

Gauquelin, M. *Cosmic influences on human behavior.* Briarcliff Manor, N.Y.: 1973.

Gelfand, S. The relationship of birth order to pain tolerance, *Journal of Clinical Psychology,* 1963, **19**, 406.

Gelfand, S., Ullmann, L. P., & Krasner, L. The placebo response: An experimental approach, *Journal of Nervous and Mental Disorders,* 1963, **136**, 379–387.

Gibson, E. J., & Walk, R. D. The "visual cliff," *Scientific American,* 1960, **202** (4), 64–71.

Goetzl, F. R. The experimental evidence for analgesic properties of antipyretic drugs, *Permanente Foundation Medical Bulletin,* 1946, **4**, 49–63.

Goldman, M. Personal communication, 1971.

Goldscheider, A. Die spezifische energie der temperaturnerven, *Monatshefte prak. Dermatol.,* 1884, **3**, 198–208, 225–241. Cited in Boring, E. G., *Sensation and perception in the history of experimental psychology.* New York: Appleton, 1942. (P. 514 ref.)

Goodall, J. van Lawick. The behavior of free-living chimpanzees in the Gombe Stream Reserve, *Animal Behavior Monographs,* 1968, **1**, Part 3.

Gottschaldt, K. The influence of past experience on the perception of figures. (1926) In Ellis, W. (Ed.), *Source book of gestalt psychology.* London: Routledge & Kegan Paul, 1938.

Greenspoon, J. The reinforcing effect of two spoken sounds on the frequency of two responses, *American Journal of Psychology,* 1955, **68**, 409–416.

Griffin, D. R. *Animal structure and function.* New York: Holt, Rinehart and Winston, 1962.

Grossman, S. P. *A textbook of physiological psychology.* New York: Wiley, 1967.

Guilford, J. P. *Psychometric methods.* New York: McGraw-Hill, 1954.

Guttman, N., & Kalish, H. I. Discriminability and stimulus generalization, *Journal of Experimental Psychology,* 1956, **51**, 79–88.

Guttman, N., & Kalish, H. I. Experiments in discrimination, *Scientific American,* 1958, **198**, 77–80.

Haas, K. Obedience to harmful orders: Do we have a phenomenon? *Psychology,* 1969, **6**, 2–5.

Haber, R. N. *Information processing approaches to visual perception.* New York: Holt, Rinehart and Winston, 1969.

Haber, R. N., & Haber, R. B. Eidetic imagery: I Frequency, *Perceptual and Motor Skills,* 1964, **19**, 131–138.

Hake, H., & Rodwan, A. S. Perception and recognition. In Sidowski, J. B. (Ed.), *Experimental methods and instrumentation in psychology.* New York: McGraw-Hill, 1966.

Hall, E. T. *The hidden dimension.* Garden City, N.Y.: Doubleday, 1966.

Hanawalt, N. G. The effect of practice upon perception of simple designs masked by complex designs, *Journal of Experimental Psychology,* 1942, **31,** 134–148.

Hardy, J. D., Wolff, B. B., & Goodell, H. *Pain sensations and reactions.* Baltimore: Williams & Wilkins, 1952.

Hashim, S. A., & Van Itallie, T. B. Studies in normal and obese subjects with a monitored food-dispensing device, *Annals of the New York Academy of Sciences,* 1965, **131,** 654–661.

Hebb, D. O. *The organization of behavior.* New York: Wiley, 1949.

Held, R. Plasticity in sensory-motor systems, *Scientific American,* 1965, **213** (5) 84–94.

Held, R. Dissociation of visual functions by deprivation and rearrangement, *Psychologische Forschung,* 1968, **31** (4), 338–348.

Held, R., & Hein, A. Movement-produced stimulation in the development of visually guided behavior, *Journal of Comparative and Physiological Psychology,* 1963, **56,** 872–876.

Helmholtz, H. V. The speed of the nerve impulse. (1852a) In Sheridan, C. L. (Ed.), *Readings for experimental psychology,* New York: Holt, Rinehart and Winston, 1972.

Helmholtz, H. V. Ueber die Theorie der Zusammenalsetzen Farben, *Annalen der Physik Una Chemie,* 1852b, **163,** 45–66.

Hering, E. *Zur lehre vom Lichtsinne.* (1874.) Cited in Boring, E. G., *Sensation and perception in the history of experimental psychology.* New York: Appleton, 1942.

Hess, E. H. Space perception in the chick, *Scientific American,* 1956, **195,** No. 1, 71–80.

Hess, E. H., & Polt, J. M. Pupil size as related to interest value of visual stimuli, *Science,* 1960, **132,** 349–350.

Hilgard, E. R. Preface to Dover Edition of Ebbinghaus' *Memory.* New York: Dover, 1964.

Hilgard, E. R. Pain as a puzzle for psychology and physiology, *American Psychologist,* 1969, **24,** 103–113.

Hilgard, E. R. A neodissociation theory of pain reduction in hypnosis, *Psychological Review,* 1973, **80,** 396–411.

Hoffman, M. *The gay world.* New York: Basic Books, 1968.

Holland, C. H. Sources of variance in the experimental investigation of behavioral obedience. Doctoral dissertation, University of Connecticut. Ann Arbor: Michigan University Microfilms, 1967, No. 69–2146.

Holz, W. C., & Azrin, N. H. Discriminative properties of punishment, *Journal of Experimental Analysis of Behavior,* **4,** 1961, 225–232.

Hooker, E. The adjustment of the overt male homosexual, *Journal of Projective Techniques,* 1957, **21,** 18–31.

Hovland, C. I. The generalization of conditioned responses I. The sensory generalization of conditioned responses with varying frequencies of tone, *Journal of General Psychology,* 1937, **17,** 125–148.

Hubel, D. H. The visual cortex of the brain, *Scientific American,* Nov. 1963, **209,** 54–62.

Hull, C. L. *Principles of behavior.* New York: Appleton, 1943.

Hull, C. L., Livingston, J. R., Rouse, V. O., & Barker, A. N. True, sham, and esophageal feeding as reinforcements, *Journal of Comparative and Physiological Psychology,* 1951, **44,** 236–245.

Jacobson, E. *You must relax.* New York: McGraw-Hill, 1957.

Jeffress, L. A. (Ed.). *Cerebral mechanisms in behavior,* The Hixon Symposium. New York: Wiley, 1951.

Johnson, L., & Naitoh, P. *The operational consequences of sleep deprivation and sleep deficit.* Neuilly sur Seine, France: Advisory Group for Aerospace Research and Development, AG–193, 1974.

Johnson, L., Naitoh, P., Moses, J., & Lubin, A. Interaction of REM deprivation and stage 4 deprivation with total sleep loss: Experiment 2, *Psychophysiology,* 1974, **11,** 147–159.

Johnson, L., Syle, E., & Dement, W. Electroencephalographic and autonomic activity during and after prolonged sleep deprivation, *Psychosomatic Medicine,* 1965, **27,** 415–423.

Jones, H., & Oswald, I. Two cases of healthy insomnia, *Electroencephlography and Clinical Neurophysiology,* 1968, **24,** 378–380.

Justesen, D., Braun, E., Garrison, R., & Pendleton, R. Pharmacological differentiation of allergic and classically conditioned asthma in the guinea pig, *Science,* 1970, **170,** 864–866. Also reprinted in Sheridan (1972).

Kalish, H. I., & Haber, A. Generalization: I. Generalization gradients from single and multiple stimulus points. II. Generalization of inhibition, *Journal of Experimental Psychology,* 1963, **65** (2), 176–181.

Kasamatsu, A., & Hirai, T. An electroencephalographic study on the Zen meditation (Zazen), *Folio Psychiatrica Neurologica Japonica,* 1966, **20,** 315–336. Also reprinted in Tart, C. (Ed.), *Altered states of consciousness.* New York: Wiley, 1969.

Keller, F. S., & Schonfeld, W. N. *Principles of psychology.* New York: Appleton, 1950.

Keppel, G. *Design and analysis: A researcher's handbook.* Englewood Cliffs, N.J.: Prentice-Hall, 1973.

Keppel, G., & Underwood, B. Proactive inhibition in short-term retention of single items, *Journal of Verbal Learning and Verbal Behavior,* 1962, **1,** 153–161.

Kiesler, C., Collins, B., & Miller, N. *Attitude change.* New York: Wiley, 1969.

Killham, W., & Mann, L. Level of destructive obedience as a function of transmittor and executant roles in the Milgram obedience paradigm, *Journal of Personality and Social Psychology,* 1974, **29,** 696–702.

King, R. G. Obedience to authority: What became of the kind-hearted woman? Unpublished M.A. thesis, University of Missouri-Kansas City, 1971.

Kintz, B., Delprato, D., Mettee, D., Persons, C., & Schaffe, R. The experimenter effect, *Psychological Bulletin,* 1965, **63,** 223–232.

Klüver, H. Eidetic phenomena, *Psychological Bulletin,* 1932, **29,** 181–203.

Koffka, K. *Principles of gestalt psychology.* New York: Harcourt, 1935.

Kohler, I. Experiments with goggles, *Scientific American,* 1962, **206** (5), 63–72.

Kohler, I. The formation and transformation of the perceptual world. (Fiss, H., Trans.) *Psychological Issues,* 1964, **3**, Entire No. 4.

Köhler, W. *The mentality of apes.* New York: Harcourt, 1925.

Köhler, W. Relational determination in perception. In Jeffress, L. A. (Ed.), *Cerebral mechanisms in behavior,* The Hixon Symposium. New York: Wiley, 1951.

Koulack, D. Rapid eye movements and visual imagery during sleep, *Psychological Bulletin,* 1972, **78**, 155–158.

Krechevsky, I. "Hypothesis" versus "chance" in the presolution period in sensory discrimination learning, *University of California Publications in Psychology,* 1932, **6**, 27–44. Reprinted in Riopelle, A. J. (Ed.), *Animal problem solving.* Baltimore: Penguin, 1967.

Krechevsky, I. A study of the continuity of the problem-solving process, *Psychological Review,* 1938, **45**, 107–133.

Kurtz, R. G., Rozin, P., & Teitlebaum, P. Ventromedial hypothalamic hyperphagia in the hypophysectomized weanling rat, *Journal of Comparative and Physiological Psychology,* 1972, **80**, 19–25.

La Piere, R. T. Attitudes vs actions, *Social Forces,* 1934, 230–237.

Lashley, K. S. *Brain mechanisms and intelligence.* Chicago: University of Chicago Press, 1929.

Lashley, K. S. The mechanism of vision: I. A method for rapid analysis of pattern-vision in the rat, *Journal of Genetic Psychology,* 1930, **37**, 453–460.

Lashley, K. S. An examination of the "continuity theory" as applied to discrimination learning, *Journal of General Psychology,* 1942, **26**, 241–265.

Lawrence, D. H. Acquired distinctiveness of cues: I. Transfer between discriminations on the basis of familiarity with the stimulus, *Journal of Experimental Psychology,* 1949, **39**, 770–784.

Lele, P. O., & Weddell, G. The relationship between neurohistology and corneal sensitivity, *Brain,* 1956, **79**, 119–154.

Lele, P. P., & Weddell, G. Sensory nerves of the cornea and cutaneous sensibility, *Experimental Neurology,* 1959, **1**, 334–359.

Lettvin, J. Y., Maturana, H. R., Pitts, W. H., & McCulloch, W. S. Two remarks on the visual system of the frog. In Rosenblith, W. A. (Ed.), *Sensory communication.* Cambridge: M.I.T. Press, 1961.

Levinson, D. M. Interocular transfer in guinea pigs following section of the corpus callosum, *Journal of Comparative and Physiological Psychology,* 1972, **78**, 26–31.

Levinson, L. *Webster's unafraid dictionary.* New York: Collier, 1967.

Levitz, L., & Stunkard, A. A therapeutic coalition for obesity: Behavior modification and patient self-help, *American Journal of Psychiatry,* 1974, **131**, 423–427.

Lindemann, H. *Relieve tension the autogenic way.* New York: Wyden, 1973.

Lindquist, E. *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin, 1953.

Lorenz, K. Z. *King Solomon's ring.* (Wilson, M. K., Trans.) New York: Crowell, 1952.

Lovejoy, E. *Attention in discrimination learning.* San Francisco: Holden-Day, 1968.

Lucas, G. The control of keypecks during automaintenance by prekeypeck omission training, *Animal Learning and Behavior,* 1975, **3**, 33–36.

Luce, G. G. *Current research on sleep and dreams.* Washington, D.C.: U. S. Department of Health, Education, and Welfare, National Institute of Mental Health, Public Health Service Publication No. 1389, 1965.

Lund, R. D. Uncrossed visual pathways of hooded and albino rats, *Science,* 1965, **149**, 1506–1507.

Luria, A. R. (Solotoraoff, L., Trans.) *The mind of a mnemonist.* New York: Basic Books, 1968.

Lynch, K., & Sheridan, C. L. Interocular transfer of a brightness and of a pattern discrimination in albino and pigmented guinea pigs, *Psychonomic Science,* 1970, **20**, 277–279.

Madjkowski, J. Electrophysiological studies of learning in split-brain cats, *Electroencephalography and clinical neurophysiology,* 1967, **23**, 521–531.

Marshall, W., Talbot, S. A., & Adis, H. W. Cortical responses of the cat to gross photic and electrical afferent stimulation, *Journal of Neurophysiology,* 1943, **6**, 1–15.

Martin, R., & Melvin, K. Vicious circle behavior as a function of delay of punishment, *Psychonomic Science,* 1964, **1**, 415–416.

Mason, J. W. Organization of psychoendocrine mechanisms: A review and reconsideration of research. In Greenfield, N. S., & Sternbach, R. A. (Eds.), *Handbook of Psychophysiology,* New York: Holt, Rinehart and Winston, 1972.

Massaro, D. *Experimental psychology and information processing.* Chicago, Rand-McNally, 1975.

Masters, W. H., & Johnson, V. E. *Human sexual response.* Boston: Little, Brown, 1966.

Mayer, J. *Overweight, causes, cost, and control.* Englewood Cliffs, N.J.: Prentice-Hall, 1968.

McCulloch, T. L., & Pratt, J. G. A study of the pre-solution period in weight discrimination by white rats, *Journal of Comparative Psychology,* 1934, **18**, 271–290.

Meddis, R., Pearson, A., & Langford, G. An extreme case of healthy insomnia. *Electroencephalography and clinical neurophysiology,* 1973, **35**, 213–214.

Meikle, T., & Sechzer, J. Interocular transfer of brightness discrimination in "split-brain" cats, *Science,* 1960, **132**, 734–735.

Meltzoff, J., & Kornreich, M. *Research in psychotherapy.* New York: Atherton Press, 1970.

Melzack, R. *The puzzle of pain.* New York: Basic Books, 1973.

Melzack, R., & Casey, K. L. Sensory, motivational and central control determinants of pain. In Kenshalo, D. R. (Ed.), *The skin senses.* Springfield, Ill.: Charles C Thomas, 1968.

Melzack, R., & Torgerson, W. On the language of pain, *Anesthesiology,* 1971, **34**, 50–59.

Melzack, R., & Wall, P. E. Pain mechanisms: A new theory, *Science,* 1965, **150**, 971–979.

Menges, R. J. Openness and honesty versus coercion and deception in psychological research, *American Psychologist,* 1973, **28**, 1030–1034.

Meyer, D. R., Treichler, F. R., & Meyer, P. M. Discrete-trial training techniques and stimulus variables. In Schrier, A. M., Harlow, H. F., & Stollnitz, F. (Eds.), *Behavior of nonhuman primates,* Vol. I. New York: Academic Press, 1965.

Milgram, S. Behavioral study of obedience, *Journal of Abnormal and Social Psychology,* 1963, **67**, 371–378.

Milgram, S. Group pressure and action against a person, *Journal of Abnormal and Social Psychology,* 1964, **69**, 137–143.

Milgram, S. Liberating effects of group pressure, *Journal of Personality and Social Psychology,* 1965a, **1**, 127–134.

Milgram, S. Some conditions of obedience and disobedience to authority, *Human Relations,* 1965b, **18**, 57–76.

Milgram, S. *Obedience to authority.* New York: Harper & Row, 1974.

Miller, A. G. Role playing: An alternative to deception? *American Psychologist,* 1972, **27**, 623–636.

Miller, G. The magical number seven, plus or minus two, *Psychological Review,* 1956, **63**, 81–97.

Miller, N., & Dworkin, B. Visceral learning: Recent difficulties with curarized rats and significant problems for human research. In Obrist, P., Black, A., Brener, J., & DiCara, L. (Eds.), *Cardiovascular psychophysiology.* Chicago: Aldine, 1974, pp. 312–331.

Mishkin, M., & Forgays, D. G. Word recognition as a function of retinal locus, *Journal of Experimental Psychology,* 1952, **43**, 43–48.

Mountcastle, V. B. (Ed.) *Medical Physiology,* 12th ed. St. Louis: Mosby, 1968.

Mueller, E. E., Loeffel, R., & Mead, S. Skin impedance in relation to pain threshold testing by electrical means, *Journal of Applied Physiology,* 1953, **5**, 746–752.

Mühl, A. Automatic writing and hypnosis. In LeCron, L. (Ed.) *Experimental hypnosis,* Secaucus, N.J.: Citadel Press, 1972, pp. 426–438.

Müller, J. *Handbuch der Physiologie,* II. Cited in Boring, E. G., *A history of experimental psychology,* 2nd ed. New York: Appleton, 1950, p. 94.

Munn, N. L. Visual pattern discrimination in the white rat, *Journal of Comparative and Physiological Psychology,* 1930, **10**, 145–166.

Munn, N. L. Concerning visual form discrimination in the white rat, *Journal of Genetic Psychology,* 1929, **36**, 291–302.

Muntz, W., & Sutherland, N. The role of crossed and uncrossed optic nerve fibers in the visual discrimination of shape by rats, *Journal of Comparative Neurology,* 1964, **122**, 69–77.

Nealy, S. M., & Edwards, B. J. "Depth perception" in rats without pattern-vision experience, *Journal of Comparative and Physiological Psychology,* 1960, **53**, 468–469.

Neisser, U. *Cognitive psychology,* New York: Appleton, 1967.

Nisbett, R. E. Hunger, obesity and the ventromedial hypothalamus, *Psychological Review,* 1972, **79**, 433–452.

Nisbett, R., & Gurwitz, S. Weight, sex, and the eating behavior of human newborns. *Journal of Comparative and Physiological Psychology,* 1970, **73**, 245–253.

Nisbett, R. E., & Schacter, S. The cognitive manipulation of pain, *Journal of Experimental Social Psychology,* 1966, **3**, 227–236.

Norman, D. *Memory and attention.* New York: Wiley, 1969.

Orne, M. T. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications, *American Psychologist,* 1962, **17,** 776–783.

Orne, M., & Holland, C. C. On the ecological validity of laboratory deceptions, *International Journal of Psychiatry,* 1968, **6,** 282–293.

Osgood, C. E. *Method and theory in experimental psychology.* New York: Oxford University Press, 1953.

Oswald, I. Number-forms and kindred visual images, *Journal of General Psychology,* 1960, **63,** 81–88.

Page, M. M. Social psychology of a classical conditioning of attitudes experiment, *Journal of Personality and Social Psychology,* 1969, **11,** 177–186.

Page, M. M. Demand awareness, subject sophistication, and the effectiveness of a verbal "reinforcement," *Journal of Personality,* 1970, **38,** 287–301.

Page, M. M. Effects of evaluation apprehension on cooperation in verbal conditioning, *Journal of Experimental Research in Personality,* 1971, **5,** 85–91.

Page, M. M. Demand characteristics and the verbal operant conditioning experiment, *Journal of Personality and Social Psychology,* 1972, **23,** 372–378.

Page, M. M. Effects of demand cues and evaluation apprehension in an attitude change experiment, *Journal of Social Psychology,* 1973, **89,** 55–62.

Page, M. M., & Lumia, A. Cooperation with demand characteristics and the bimodal distribution of verbal conditioning data, *Psychonomic Science,* 1968, **12,** 243–244.

Paivio, A. *Imagery and verbal processes.* New York: Holt, Rinehart and Winston, 1971.

Paivio, A., Smythe, P., and Yuille, J. Imagery versus meaningfulness of nouns in paired-associate learning, *Canadian Journal of Psychology,* 1968, **22,** 427–441.

Pastore, R., & Scheirer, C. Signal detection theory: Considerations for general application, *Psychological Bulletin,* 1974, **81,** 945–958.

Patty, R., & Page, M. Manipulations of a verbal conditioning situation based upon demand characteristics theory, *Journal of Experimental Research in Personality,* 1973, **6,** 307–313.

Pavlov, I. P. *Lectures on conditioned reflexes* (Gantt, W. H., Trans.). New York: International Universities, 1928.

Perls, F., Hefferline, R., & Goodman, R. *Gestalt therapy.* New York: Julian Press, 1951.

Peterson, L., & Peterson, M. Short-term retention of individual verbal items, *Journal of Experimental Psychology,* 1959, **58,** 193–198.

Petre, C. C., & Sheridan, C. L. Interocular transfer in albino and pigmented perinatal guinea pigs, *Psychonomic Science,* 1966, **6,** 215–216.

Petrie, A. *Individuality in pain and suffering.* Chicago: University of Chicago Press. 1967.

Pfister, H. Ueber das Verhalten der Huhner beim tragen von prismen. Unpublished doctoral dissertation, University of Innsbruck, 1955. Cited in

Smith, K. U., & Smith, W. M., *Perception and motion.* Philadelphia: Saunders, 1962.

Pliner, P. Effect of liquid and solid preloads on eating behavior of obese and normal persons, *Physiology and Behavior,* 1973, **11,** 286–290.

Poser, E. G. A simple and reliable apparatus for the measurement of pain, *American Journal of Psychology,* 1962, **75,** 304–305.

Postman, L. Perception and learning. In Koch, S. (Ed.), *Psychology, a study of a science,* Vol. 5. New York: McGraw-Hill, 1963.

Poulton, E. The new psychophysics: Six models for magnitude estimation, *Psychological Bulletin,* 1968, **69,** 1–19.

Price, R. H. Signal detection methods in personality and perception, *Psychological Bulletin,* 1966, **66,** 55–62.

Pyles, M. K., Stalz, H. R., & Macfarlane, J. W. The accuracy of mothers: Reports on birth and developmental data, *Child Development,* 1935, **6,** 165–176.

Reid, L. S. Development of non-continuity behavior through continuity learning, *Journal of Experimental Psychology,* 1953, **46,** 107–112.

Reynolds, R. An irritative hypothesis concerning the hypothalamic regulation of food intake, *Psychological Review,* 1965, **72,** 105–116.

Riesen, A. H. Arrested vision, *Scientific American,* 1950, **183,** 16–19.

Riesen, A. H. Stimulation as a requirement for growth and function in behavioral development. In Fiske, D. W., & Maddi, S. R. (Eds.), *Functions of varied experience.* Homewood, Ill.: Dorsey Press, 1961.

Robbins, L. C. The accuracy of parental recall of aspects of child rearing practices, *Journal of Abnormal and Social Psychology,* 1963, **66,** 261–270.

Robbins, M. J., & Meyer, D. R. Motivational control of retrograde amnesia, *Journal of Experimental Psychology,* 1970, **84,** 220–225.

Rokeach, M. *The open and closed mind.* New York: Basic Books, 1960.

Rosenberg, M. J. When dissonance fails: On eliminating evaluation apprehension from attitude measurement, *Journal of Personality and Social Psychology,* 1965, **1,** 28–42.

Rosenthal, R., & Lawson, R. A longitudinal study of the effects of experimenter bias on the operant learning of laboratory rats, *Journal of Psychiatric Research,* 1964, **2** (2), 61–72.

Rosensweig, S. The experimental situation as a psychological problem, *Psychological Review,* 1933, **40,** 337–354.

Rossi, P. J. Adaptation and negative aftereffect to lateral optical displacement in newly hatched chicks, *Science,* 1968, **160,** 430–431.

Rozin, P., & Kalat, J. Specific hungers and poison avoidance as adaptive specializations of learning, *Psychological Review,* 1971, **78,** 459–486.

Rundus, D., Loftus, G., & Atkinson, R. Immediate free recall and three-week delayed recognition, *Journal of Verbal Learning and Verbal Behavior,* 1970, **9,** 684–688.

Sanderson, K. J., Kaas, J. H., & Guillery, R. W. Abnormal retinogeniculate and geniculocortical projections in the albino allelomorphic series of mammals, from Program and abstracts, Society for Neuroscience, 1972, p. 145.

Schachter, S. Cognitive effects on bodily functioning: Studies of obesity and eating. In Glass, D. C. (Ed.), *Neurophysiology and emotion.* New York: Rockefeller University Press, 1967.

Schachter, S. Some extraordinary facts about obese humans and rats, *American Psychologist,* 1971, **26**, 129–144.
Schachter, S., & Singer, J. Cognitive, social, and physiological determinants of emotional state, *Psychological Review,* 1962, **69**, 379–399.
Schiller, P. H. Innate constituents of complex responses in primates, *Psychological Review,* 1952, **59**, 177–191.
Schwartz, C. B. Visual discrimination of camouflaged figures. Unpublished doctoral dissertation, University of California, Berkeley, 1961.
Sechzer, J. Successful interocular transfer of pattern discrimination in "split-brain" cats with shock avoidance motivation, *Journal of Comparative and Physiological Psychology,* 1964, **58**, 76–83.
Seeman, W., & Marks, P. The behavior of the psychologist at a choice point, *American Scientist,* 1962, **50**, 538–547. Also reprinted in Sheridan, 1972.
Seligman, M. On the generality of the laws of learning, *Psychological Review,* 1970, **77**, 406–418.
Sheffield, F. D., & Roby, T. B. Reward value of a nonnutritive sweet taste, *Journal of Comparative and Physiological Psychology,* 1950, **43**, 471–481.
Sheridan, C. L. Interocular transfer of brightness and pattern discriminations in normal and corpus callosum sectioned rats, *Journal of Comparative and Physiological Psychology,* 1965a, **59**, 292–294.
Sheridan, C. L. Interocular interaction of conflicting discrimination habits in the albino rat: A preliminary report, *Psychonomic Science,* 1965b, **3**, 303–304.
Sheridan, C. L. *Fundamentals of experimental psychology.* New York: Holt, Rinehart and Winston, 1971.
Sheridan, C. L. *Readings for experimental psychology.* New York: Holt, Rinehart and Winston, 1972.
Sheridan, C. L., & King, R. G. Obedience to authority with an authentic victim, *Proceedings of the American Psychological Association,* 1972, **7**, 165–166.
Sheridan, C. L., & Shrout, L. L. Interocular transfer in the rat: The role of the occlusion process, *Psychonomic Science,* 1965, **4**, 177–178.
Sidman, Murray. *Tactics in scientific research.* New York: Basic Books, 1960.
Siegel, S. *Nonparametric statistics.* New York: McGraw-Hill, 1956.
Siipola, E., & Hayden, S. Exploring eidetic imagery among the retarded, *Perceptual and Motor Skills,* 1965, **21**, 275–286.
Silverman, I., & Regula, C. Evaluation apprehension, demand characteristics, and the effects of distraction on persuasability, *Journal of Social Psychology,* 1968, **75**, 273–281.
Silverman, I., & Shulman, A. A conceptual model of artifact in attitude change studies, *Sociometry,* 1970, **33**, 97–107.
Singh, D. Preference for mode of obtaining reinforcement in rats with lesions in septal or ventromedial hypothalamic area, *Journal of Comparative and Physiological Psychology,* 1972, **80**, 259–268.
Singh, D. Role of response habits and cognitive factors in determination of behavior of obese humans, *Journal of Personality and Social Psychology,* 1973, **27** (No. 2), 220–238.
Singh, D., & Sikes, S. Role of past experience on food-motivated behavior of obese humans, *Journal of Comparative and Physiological Psychology,* 1974, **86**, 503–508.

Singh, D., Swanson, J., Letz, R., & Sanders, M. Performance of obese humans on transfer of training and reaction time tests, *Psychosomatic Medicine,* 1973, **35,** 240–249.

Siprelle, C. Induced anxiety, *Psychotherapy: Theory, research and practice,* 1967, **4,** 36–40.

Skinner, B. F. *The behavior of organisms.* New York: Appleton, 1938.

Skinner, B. F. *Science and human behavior.* New York: Appleton, 1953.

Skinner, B. F. *Verbal behavior.* New York: Appleton, 1957.

Smith, C. S. Matter versus materials: A historical view, *Science,* 1968, **162,** 637–644.

Smith, K. U., & Smith, W. M. *Perception and motion.* Philadelphia: Saunders, 1962.

Spence, K. W. The nature of discrimination learning in animals, *Psychological Review,* 1936, **43,** 427–449.

Sperling, G. The information available in brief visual presentations, *Psychological Monographs,* 1960, **74** (Whole No. 498).

Sperry, R. W. Mechanisms of neural maturation. In Stevens, S. S. (Ed.), *Handbook of experimental psychology.* New York: Wiley, 1951.

Sperry, R. W. Hemisphere disconnection and unity in conscious awareness, *American Psychologist,* 1968, **23,** 723–733. Also reprinted in Sheridan, 1972.

Stagner, R., & Solley, C. M. *Basic psychology.* New York: McGraw-Hill, 1970.

Stass, J. W., & Willis, F. N. Eye contact, pupil dilation, and personal preference, *Psychonomic Science,* 1967, **7,** 375–376.

Staub, E. Helping a distressed person: Social personality and stimulus determinants. In Berkowitz, L. (Ed.) *Advances in experimental social psychology,* Vol. 7. New York: Academic Press, 1974.

Sternbach, R. *Pain.* New York: Academic Press, 1968.

Stevens, J. C., Herrnstein, R., & Reynolds, G. *Laboratory experiments in psychology,* New York: Holt, Rinehart and Winston, 1965.

Stevens, S. S. The psychophysics of sensory functions. In Rosenblith, W. A., *Sensory communication.* Cambridge: M.I.T. Press, 1961.

Stromeyer, C., & Psotka, J. The detailed texture of eidetic images, *Nature,* 1970, **225,** 346–349.

Stunkard, A. Obesity and the denial of hunger, *Psychosomatic Medicine,* 1959, **21,** 281–289.

Stunkard, A. Hunger and satiety, *American Journal of Psychiatry,* 1961, **118,** 212–217.

Stunkard, A. New therapies for the eating disorders, *Archives of General Psychiatry,* 1972, **26,** 1972.

Stunkard, A., & Mendelson, M. Obesity and the body image: I. Characteristics of disturbances in the body image of some obese persons, *American Journal of Psychiatry,* 1967, **123,** 1296–1300.

Sullivan, B. Effects of progressive relaxation on verbal, behavioral, and physiological measures of ischemic pain. Unpublished M.A. thesis, University of Missouri—Kansas City, 1973.

Sutcliffe, J. P. "Credulous" and "skeptical" views of hypnotic phenomena; experiments on esthesia hallucination, and delusion, *Journal of Abnormal and Social Psychology,* 1961, **62,** 189–200.

Swets, J. A., Tanner, W. P., & Birdsall, T. G. Decision processes in perception, *Psychological Review,* 1961, **68**, 301–320.

Taylor, D. W., Garner, W. R., & Hunt, H. F. Education for research in psychology, *American Psychologist,* 1959, **14**, 167–179.

Terman, L. M., & Oden, M. H. *The gifted group at midlife.* Stanford, Calif.: Stanford University Press, 1959.

Terrace, H. Discrimination learning with and without errors, *Journal of the Experimental Analysis of Behavior,* 1963, **6**, 1-27.

Teuber, H. L. Perception after brain injury. In Eccles, J. C. (Ed.), *The brain and conscious experience.* New York: Springer-Verlag, 1965.

Thorndike, E. L. Animal intelligence, an experimental study of the associative process in animals, *Psychological Monographs,* 1898, **2** (Whole No. 8).

Tilker, H. A. Socially responsible behavior as a function of observer responsibility and victim feedback, *Journal of Personality and Social Psychology,* 1970, **14**, 95–100.

Uhlela, Z., & Adams, D. Detection theory and expectations for social reinforcers: An application to aggression, *Psychological Review,* 1973, **80**, 439–445.

Underwood, B., & Schulz, R. W. *Meaningfulness and verbal learning.* Philadelphia: Lippincott, 1960.

Valentine, C. W. The innate bases of fear, *Journal of Genetic Psychology,* 1930, **37**, 394–419.

Walk, R. D. A study of visual depth and distance perception in animals. In Lehrman, D. S., Hinde, R. A., & Shaw, E. *Advances in the study of behavior.* New York: Academic Press, 1965.

Walk, R. D., & Gibson, E. J. A comparative and analytical study of visual depth perception, *Psychological Monographs,* 1961, **75** (No. 519).

Watson, J. B., & Rayner, R. Conditioned emotional reaction, *Journal of Experimental Psychology,* 120, **3**, 1–14.

Weinstein, S. Intensive and extensive aspects of tactile sensitivity as a function of body part, sex, and laterality. In Kenshalo, D. R. (Ed.), *The skin senses.* Springfield, Ill.: Charles C Thomas, 1968, 195–222.

Weiss, P. $1 + 1 \neq 2$ (when one plus one does not equal two). In Quarton, G. C., Melnechuk, T., & Schmitt, F. O., *The neurosciences.* New York: Rockefeller University Press, 1967.

Wertheimer, M. Principles of perceptual organization. (1923) In Beardslee, D. C., & Wertheimer, M., *Readings in perception.* Princeton, N.J.: Van Nostrand, 1958.

West, J., & Toonder, J. *The case for astrology.* Baltimore: Penguin, 1973.

Wickelgren, W. The long and the short of memory, *Psychological Bulletin,* 1973, **80**, 425–438.

Wickens, D. D., & Meyer, D. R. *Psychology* (rev. ed.). New York: Holt, Rinehart & Winston, 1961.

Wickens, D. D., Schroder, H. M., & Snide, J. D. Primary stimulus generalization of the GSR under two conditions, *Journal of Experimental Psychology,* 1954, **47**, 52–56.

Wiesel, T., & Hubel, D. Extent of recovery from the effects of visual deprivation in kittens, *Journal of Neurophysiology,* 1965, **28**, 1060–1072.

Wike, E. *Data analysis.* Chicago and New York: Aldine-Atherton, 1971.

Wilcoxon, H., Dragoin, W., & Kral, P. Illness-induced aversions in rat and quail: Relative salience of visual and gustatory cues, *Science,* 1971, **171,** 826–828.

Williams, H. L., Hammack, J. T., Daly, R. L., Dement, W. C., & Lubin, A. Responses to auditory stimulation, sleep loss, and the EEG stages of sleep, *EEG Journal of Clinical Neurophysiology,* 1964, **16,** 269–279.

Williams, H., & Williams, C. Nocturnal EEG profiles and performance, *Psychophysiology,* 1966, **3,** 164–175.

Willis, F. N., Initial speaking distance as a function of the speaker's relationship, *Psychonomic Science,* 1966, **5,** 221–222.

Willis, R. H., & Willis, Y. A., Role playing versus deception: An experimental comparison, *Journal of Personality and Social Psychology,* 1970, **60,** 474–477.

Wilson, E. Bright. *An introduction to scientific research.* New York: McGraw-Hill, 1952.

Woodford, F. P. (Ed.) *Scientific writing for graduate students.* New York: Rockefeller University Press, 1968.

Wooley, O., Wooley, S., & Dunham, R. Can calories be perceived and do they affect hunger in obese and nonobese humans, *Journal of Comparative and Physiological Psychology,* 1972a, **50,** 250–258.

Wooley, O, Wooley, S., & Dunham, R. Calories and sweet taste: Effects on sucrose preference in the obese and nonobese, *Physiology and Behavior,* 1972b, **9,** 765–768.

Wooley, S. Physiologic versus cognitive factors in short term food regulation in the obese and nonobese, *Psychosomatic Medicine,* 1972, 34, 62–68.

Yates, F. *The art of memory.* London: Routledge & Kegan Paul, 1966.

Young, J. Z. *Doubt and certainty in science.* Oxford: Clarendon Press, 1951.

INDEXES

Name Index

Subject Index